Global genome diversity of the *Leishmania donovani* complex

Susanne U. Franssen[1]*, Caroline Durrant[1], Olivia Stark[2], Bettina Moser[2], Tim Downing[1,3], Hideo Imamura[4], Jean-Claude Dujardin[4,5], Mandy Sanders[1], Isabel Mauricio[6], Michael A. Miles[7], Lionel F. Schnur[8], Charles L. Jaffe[8], Abdelmajeed Nasereddin[8], Henk Schallig[9], Matthew Yeo[7], Tapan Bhattacharyya[7], Mohammad Z. Alam[10], Matthew Berriman[1], Thierry Wirth[11,12]*, Gabriele Schönian[2]*, James A. Cotton[1]*

[1] Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom.
[2] Charité Universitätsmedizin, Berlin, Germany.
[3] Dublin City University, Dublin, Ireland.
[4] Institute of Tropical Medicine, Antwerp, Belgium.
[5] Department of Biomedical Sciences, University of Antwerp, Belgium.
[6] Universidade Nova de Lisboa Instituto de Higiene e Medicina, Lisboa, Portugal.
[7] London School of Hygiene and Tropical Medicine, London, United Kingdom.
[8] Kuvin Centre for the Study of Infectious and Tropical Diseases, IMRIC, Hebrew University-Hadassah, Medical School, Jerusalem, Israel.
[9] Amsterdam University Medical Centres – Academic Medical Centre at the University of Amsterdam, Department of Medical Microbiology – Experimental Parasitology, Amsterdam, The Netherlands.
[10] Department of Parasitology, Bangladesh Agricultural University, Mymensingh 2202, Bangladesh.
[11] Institut de Systématique, Evolution, Biodiversité, ISYEB, Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France.
[12] École Pratique des Hautes Études, PSL, Paris, France.

*authors for correspondence

Abstract

Protozoan parasites of the *Leishmania donovani* complex – *L. donovani* and *L. infantum* – cause the fatal disease visceral leishmaniasis. We present the first comprehensive genome-wide global study, with 151 cultured field isolates representing most of the geographical distribution. *L. donovani* isolates separated into five groups that largely coincide with geographical origin but vary greatly in diversity. In contrast, the majority of *L. infantum* samples fell into one globally-distributed group with little diversity. This picture is complicated by several hybrid lineages. Identified genetic groups vary in heterozygosity and levels of linkage, suggesting different recombination histories. We characterise chromosome-specific patterns of aneuploidy and identified extensive structural variation, including known and suspected drug resistance loci. This study reveals greater genetic diversity than suggested by geographically-focused studies, provides a resource of genomic variation for future work and sets the scene for a new understanding of the evolution and genetics of the *Leishmania donovani* complex.

Introduction

The genus *Leishmania* is a group of more than 20 species of protozoan parasites that cause the neglected tropical disease leishmaniasis in humans, but also infect other mammalian hosts. Leishmaniasis is transmitted by phlebotomine sandflies and exists in four main clinical conditions: cutaneous leishmaniasis (CL), seen as single and multiple cutaneous lesions; mucocutaneous leishmaniasis (MCL), presenting in mucosal tissue; diffuse cutaneous leishmaniasis (DCL), seen as multiple nodular cutaneous lesions covering much of the body; and visceral leishmaniasis (VL, also known as kala-azar), affecting internal organs. Disease prevalence is estimated at 0.9 to 1.6 million new cases, mostly of CL, and up to 90,000 new cases per year of VL are associated with a 10% mortality rate (Alvar et al., 2012; Burza et al., 2018). The form of the disease is largely driven by the species of *Leishmania* causing the infection but is further influenced by vector biology and host factors, importantly by host immune status (Burza et al., 2018; McCall et al., 2013). In the mammalian host, parasites are intracellular, residing mainly in long lived macrophages. In the most severe visceral form, parasites infect the spleen, liver, bone marrow and lymph nodes, leading to splenomegaly and hepatomegaly. This results in a range of symptoms including frequent anaemia, thrombocytopenia and neutropenia, and common secondary infections which are often fatal without successful treatment (for review see: Rodrigues *et al.*, 2016; Burza *et al.*, 2018), although most infections remain asymptomatic (Ostyn et al., 2011).

The key species responsible for VL are *L. donovani* and *L. infantum* (see reviews McCall *et al.*, 2013; Burza *et al.*, 2018), which together form the *L. donovani* species complex. Both species mainly cause VL, but for each species atypical cutaneous presentations are common in some foci (reviewed in Thakur *et al.* 2018; e.g. Guerbouj *et al.*, 2001; Zhang *et al.*, 2014). Post-kala-azar dermal leishmaniasis (PKDL), is a common sequel to VL that manifests with dermatological symptoms appearing after apparent cure of the visceral infection. PKDL is mainly seen on the Indian subcontinent and north-eastern and eastern Africa following infections caused by *L. donovani* (Zijlstra et al., 2003). *L. donovani* is considered to be largely anthroponotic even if the parasites can be encountered in animals (Bhattarai et al., 2010). In contrast, *L. infantum* – like most *Leishmania* species – causes a zoonotic disease, where dogs are the major domestic reservoir but a range of wild mammals can also be involved in transmission (Díaz-Sáez et al., 2014; Quinnell and Courtenay, 2009). Both species are widespread across the globe, with major foci in the Indian subcontinent and East Africa for *L. donovani*, the Mediterranean region and the Middle East for *L. infantum*, and China for both species (Lun et al., 2015; Lysenko, 1971; Ready, 2014). *L. infantum* has also more recently spread to the New

World, via European migration during the 15th or 16th Century (Leblois et al., 2011), where it was sometimes described as a third species, *L. chagasi*. Leishmaniasis caused by parasites of the *L. donovani* complex differs across and even within geographical locations in the nature and severity of clinical symptoms (e.g. Guerbouj *et al.*, 2001; Zhang *et al.*, 2014; Thakur *et al.*, 2018) and in the species of phlebotomine sandflies that act as vectors (Alemayehu and Alemayehu, 2017).

For this important human pathogen, there is a long history of interest in many aspects of the basic biology of *Leishmania*, including extensive interest in epidemiology, cell biology and immunology as well as the genetics and evolution of these parasites (e.g. Simpson *et al.*, 2006; Quinnell and Courtenay, 2009; Mougneau *et al.*, 2011). *Leishmania* has two unusual genomic features that influence its genetics, including mosaic aneuploidy and a complex and predominantly clonal life cycle. Aneuploidy is the phenomenon where individual chromosomes within a cell are of different copy numbers, and mosaic aneuploidy is where the pattern of chromosome dosage varies between cells of a clonal population (Bastien et al., 1990; Sterkers et al., 2011). Genome sequencing studies have shown extensive aneuploidy in cultured *Leishmania* field isolates (e.g. Downing *et al.*, 2011; Rogers *et al.*, 2014; Zhang *et al.*, 2014; Imamura *et al.*, 2016). Variation in chromosome dosage appears to be greater in *in vitro* than *in vivo* in animal models (Dumetz et al., 2017) or human tissues (Domagalska et al., 2019). However, these studies estimate average dosage of chromosomes in a population of sequenced cells. Only a few studies have directly investigated mosaicism between cells and these found it to be extensive both *in vitro* (Sterkers *et al.*, 2011; Lachaud *et al.*, 2014) and *in vivo* (Barja et al., 2017). Reproduction was originally thought to be predominantly clonal and this is still assumed to be the only mode of reproduction for the intracellular amastigotes found in the mammalian host. A number of studies have shown that hybridisation can occur during passage in the sandfly vector. This was demonstrated experimentally (e.g. Akopyants *et al.*, 2009; Romano *et al.*, 2014; Inbar *et al.*, 2019) also showing evidence of meiosis (Inbar et al., 2019) and in field isolates through recombination-like signatures (Cotton et al., 2019; Rogers et al., 2014). However, the incidence of sexual reproduction in natural populations is still unclear (Ramírez and Llewellyn, 2014).

Despite this research, much remains unclear about the diversity, evolution and genetics of the *L. donovani* species complex. Difficult and laborious isoenzyme typing (Rioux et al., 1990) dominated the description of *Leishmania* populations for at least 25 years (Schönian et al., 2011) but suffered from a critical lack of resolution, leading to convergent signals (Jamjoom et al., 2004). More recent typing schemes, using variation at small numbers of genetic loci (multi-locus sequence typing, MLST) or microsatellite repeats (multi locus microsatellite typing, MLMT) improved the resolution of

Leishmania phylogenies and enabled population genetic analyses (Gouzelou et al., 2012; Herrera et al., 2017; Kuhls et al., 2007; Schönian et al., 2011) but are hard to compare when using different marker sets (Schönian et al., 2011). In contrast, genome-wide polymorphism data offers much greater resolution (Downing et al., 2011; Rogers et al., 2014), provides richer information on aneuploidy and other classes of variants , i.e. SNPs, small indels and structural variants, and enables insights into gene function from genome-wide studies of selection and association mapping (Carnielli et al., 2018; Downing et al., 2011). Moreover, advances in DNA sequencing technology together with the availability of reference genome assemblies for most of the clinically important species (Downing et al., 2011; González-de la Fuente et al., 2019; Peacock et al., 2007; Real et al., 2013; Rogers et al., 2011) in freely accessible public databases (Aslett et al., 2010) now make it feasible to sequence collections of isolates and determine genetic variants genome-wide. Several studies on the *L. donovani* complex have applied such an approach including foci in Nepal (16 isolates, Downing *et al.*, 2011), Turkey (12 isolates, Rogers *et al.*, 2014), the Indian subcontinent (204 isolates, Imamura *et al.*, 2016), Ethiopia (41 isolates from 16 patients, Zackay *et al.*, 2018) and Brazil (20 and 26 isolates, respectively, Teixeira *et al.*, 2017; Carnielli *et al.*, 2018). However, genomic studies to date have addressed genome-wide diversity in geographically restricted regions, leaving global genome diversity in the species complex unknown.

We present whole-genome sequence data from isolates of the *L. donovani* species complex across its global distribution. Our genome-wide SNP data revealed the broad population structure of the globally distributed samples from the species complex. *L. infantum* samples from across the sampling range fall mainly into a single clade, while *L. donovani* is much more diverse, largely reflecting the geographical distribution of the parasites. As expected, parasites from the New World appeared closely related to parasites found in Mediterranean Europe. In addition to SNP diversity, we identified characteristic aneuploidy patterns of *in vitro* isolates shared across populations, variable heterozygosity between groups, differing levels of within-group linkage suggesting different recombination histories within geographical groups, and extensive structural diversity. This analysis reveals a much greater genetic diversity than suggested by previous, geographically-focused whole-genome studies in *Leishmania* and sets the scene for a new understanding of evolution in the *Leishmania donovani* species complex.

Results

Whole-genome variation data of 151 isolates of the *L. donovani* complex

We generated paired-end Illumina whole-genome sequence data from promastigote cultures of 98 isolates from the *L. donovani* complex. These sequence data resulted in a median haploid genome coverage ranging between 10 and 88 (median=27) when mapped against the reference genome assembly of *L. infantum* JPCM5 (MCAN/ES/98/LLM-724; Peacock et al., 2007). These data were combined with previously published sequence data for parasites from Turkey (N=11; Rogers *et al.*, 2014), Sri Lanka (N=2; Zhang *et al.*, 2014), Spain (N=1; Peacock *et al.*, 2007), Ethiopia (N=6; Zackay *et al.*, 2018) and a subset of the extensive dataset available from the Indian subcontinent (N=33; Imamura *et al.*, 2016) resulting in a total of 151 isolates (Table S1, visualised at https://microreact.org/project/_FWlYSTGf; Argimón et al., 2016).

Accurate SNP variants were identified with a genotype quality of at least 10 (median=99), indicating a <0.1 (median=~10^{-10}) probability of an incorrect genotype call across 87.8% of the reference genome. The remaining 12.2% could not be assayed as short reads could not be uniquely mapped to repetitive parts of the genome. This identified a total of 395,624 SNP sites out of the ~32 Mb reference assembly. We also used these sequence data to infer extensive gene copy-number variation (91.5% of genes varied in dosage; 7,625 / 8,330 genes) and larger genome structure variation, including copy numbers of individual chromosomes (aneuploidy) that is common in *Leishmania*. Together, these data represent the most comprehensive, global database of genetic variation available for any *Leishmania* species.

Evolution of the *L. donovani* complex

Phylogenetic reconstruction based on whole-genome SNP variation clearly separated *L. infantum* from *L. donovani* strains. *L. donovani* separated into five major groups that coincide with geographic origin (Fig. 1 A-B, S1). While the inferred root of the phylogeny is between *L. infantum* and *L. donovani*, groups within *L. donovani* showed similar levels of divergence as between the two species, with the deepest branches within *L. donovani* in East Africa. The largest *L. donovani* group in our collection, Ldon1, included samples from the Indian subcontinent, and could be further divided into two subgroups that separate samples from India, Nepal and Bangladesh from three samples of Sri Lankan origin; both subgroups displayed strikingly little diversity. The large number of isolates in Ldon1 is due

to the extensive previous genomic work in this population (Downing et al., 2011; Imamura et al., 2016), which identified this as the 'core group' of strains circulating in the Indian subcontinent. The genetically and geographically closest group, Ldon2, was restricted to the Nepali highlands and also includes the more divergent sample, BPK512A1 (Lon2 is the ISC1 group of Imamura *et al.*, 2016). The latter isolate shared sequence similarity with a far more diverse group, Ldon4, of parasites from the middle East (Iraq and Saudi Arabia) and Ethiopia (Fig. 1 A). Admixture analysis identified three additional samples (from Sudan and Israel), to be of mixed origin between groups Ldon3 and Ldon4. The Ldon3 group is restricted to Sudan and northern Ethiopia and an outlier sampled in Malta likely represents an imported case. Group Ldon5 displayed little diversity and is mainly confined to Southern Ethiopia and Kenya, with the rift valley in Ethiopia presumably restricting genetic exchange with Ldon3 through different sandfly vectors (Gebre-Michael et al., 2010; Gebre-Michael and Lane, 1996). A single outlier from this group, LRC-L51p, was sampled in India and again presumably represents an imported case of African origin.

In contrast, most of the samples of *L. infantum* clustered into a single group, Linf1, with relatively little diversity within the group but a broad geographical distribution including China, Central Asia, the Mediterranean Region and Latin America (Fig. 1 A-B). Admixture analysis using different numbers of total populations (*K*) divided this group into two to three subgroups, separating samples from China, Uzbekistan and a single Israeli isolate, from two groups that both include samples from South America and the Mediterranean region. The latter two subgroups correspond to MON1 (31 samples of the largest subgroup; Fig. 1 A, S1) and non-MON1 zymodemes (6 samples from Europe, Turkey and Panama) typed by Multilocus Enzyme Electrophoresis (MLEE) (Rioux et al., 1990). Therefore, geography is not the main driver of parasite diversity in *L. infantum*. In contrast to the low diversity across the wide geographical range of the core *L. infantum* group, the remaining samples of *L. infantum*, from Cyprus and Çukurova in Turkey, are genetically more distinct and showed unusual positioning in the phylogeny close to the split between *L. infantum* and *L. donovani*. Samples from the Çukurova region of Turkey (CUK, green) are considered to be a lineage descended from a single crossing event of a strain related to the *L. infantum* reference strain JPCM5 and an unknown *L. infantum* or *L. donovani* strain (Rogers et al., 2014). Isolates from Cyprus (CH, grey) are also divergent from the L. *infantum* group: these parasites were identified as *L. donovani* using MLEE, but the associated pattern of markers (MON-37) has been shown to be paraphyletic (Alam et al., 2009), so its species identity might be debateable. Our data suggest that the two slightly different Cypriot isolates (CH32 and CH34) are admixed between the Çukurova and remaining Cypriot strains. Two more isolates

(MAM and EP; from Brazil and Turkey) are both highly divergent from any other isolates in the phylogeny, and appeared to be admixed between the Linf1 group and other lineages.

Aneuploidy

We observed extensive variation in chromosome copy number in our isolated strains *in vitro*, inferred from read coverage depth, with the pattern of variation being incongruent with the genome-wide phylogeny (Fig. S2). Aneuploidy patterns are known to vary over very short time scales, even within strains and upon changing environments (Dumetz et al., 2017; Lachaud et al., 2014; Sterkers et al., 2011), although consistent patterns of aneuploidy have been observed within small groups of closely related cultured field isolates (Imamura et al., 2016). We took advantage of the greater diversity and global scope of our data to investigate somy patterns of cultivated promastigotes for individual chromosomes across geographically distinct groups. As expected, the majority of chromosomes had a median somy of two across isolates, apart from chromosomes 8, 9 and 23 and chromosome 31 with a median somy of three and four, respectively (Fig. S3 A). However, trisomy was widespread with all chromosomes being overall trisomic in at least two isolates (2%) and at least half of all chromosomes were trisomic in ≥28 isolates (19%). In contrast, monosomy was rare – with only four chromosomes having somy of one in a single isolate each. As previously reported for *Leishmania* (e.g. Akopyants *et al.*, 2009; Downing *et al.*, 2011; Imamura *et al.*, 2016), chromosome 31 was unusual in being dominantly tetrasomic (81% of samples) and we observed no somy levels below three. Much of this pattern – general disomy, with occasional trisomy and sporadic higher dosage for most chromosomes – was consistent across the four largest groups, as was the high dosage of chromosome 31 (Fig. S3 B). Similarly, chromosome 23 showed a tendency to trisomy in all four groups, and chromosomes 8 and 9 were dominantly trisomic in three of the groups. Other chromosomes, including 5, 6, 7, 13, 26 and 35, showed different patterns of dominant somies between groups (Fig. 2 A, S3 B).

As some chromosomes appeared to be more frequently present at high copy numbers in our isolates, we investigated whether their copy numbers were also more variable. Copy number variability for each chromosome was estimated by the standard deviation (sd) in somy and was positively correlated between the four largest groups (Fig. 2 B). Correlations were much higher between three groups from diverse sampling locations, while correlations to CUK group sampled in the Çukurova province were lower, suggesting a distinct pattern of aneuploidy variability in this group – perhaps due to its hybrid origin (Rogers et al., 2014). Given the positive correlations between independent groups, we investigated chromosome-specific variation in somy using the four independent groups (Fig. 2 C). A

few chromosomes including 19, 27, 28 and 34 showed almost no variation, while several chromosomes showed very high variation in chromosome copy number with the top five chromosomes being 23, 5, 8, 6 and 26 (Fig. 2 C). This indicated that some chromosomes have higher propensities for chromosome aneuploidy turnover than others.

Heterozygosity

Samples varied greatly in genome-wide heterozygosity: 70% of the isolates in our collection showed extremely low heterozygosity (< 0.004; see Material & Methods) corresponding to 23 - 2,057 (median=80) heterozygous sites per sample. The remaining high-heterozygosity samples largely showed heterozygosities up to ~0.02 (equivalent to 15,281 heterozygous sites per sample) with a few outliers exceeding this threshold and reaching a heterozygosity of 0.065 in one isolate (MAM, 50,543 heterozygous sites) (Fig. 3 A). For almost all isolates the majority of genome-wide 10 kb windows had almost no heterozygous sites: only 11 isolates had a median count of heterozygous sites per window greater than zero (Fig. S4). This predominant homozygosity for the majority of isolates of the *L. donovani* complex was in striking contrast to expectations for sexual populations under Hardy-Weinberg equilibrium, or for clonally reproducing populations: clonal reproduction is expected to increase heterozygosity, as single mutations cannot be assorted to form novel homozygous genotypes (Balloux et al., 2003; De Meeûs et al., 2006; Weir et al., 2016). Most main groups were dominated by samples of low heterozygosity, with the exception of the Ldon3 group and the CUK group of hybrid *L. infantum* isolates (Rogers et al., 2014). Other high-heterozygosity isolates mainly appeared in positions intermediate between large groups in the phylogeny, and showed mixed ancestry in the admixture analysis (e.g. isolates MAM, EP, CH32, CH34, GE, LEM3472, LRC-L740; Fig. 1 A), leading us to hypothesise that they represent recent hybrids between the distinct, well-differentiated populations.

The low heterozygosity together with strong genetic signatures of inbreeding in *Leishmania* had previously been identified using MLST and microsatellite data, and has generally been attributed to extensive selfing between cells from the same clone (Ramírez and Llewellyn, 2014; Rougeron et al., 2009). However, an alternative explanation could be that frequent aneuploidy turnover also reduces within-cell heterozygosity if an alternate haplotype is lost during somy reduction (Sterkers et al., 2014). We therefore tested whether the chromosome-specific variation in somy for each group was negatively correlated with chromosome-specific sample heterozygosity, as a high turnover rate could reduce within-strain heterozygosity. Linear regressions for the different groups showed negative

slopes for three of seven groups but only the slope for the Ldon3 group was significant after multiple testing correction (Fig. 3 B). For the four groups, Ldon1, Ldon2, Ldon5 and Linf1, where the regression slope was almost zero, the chromosomes were almost completely homozygous which might make potential effects undetectable (Fig. 3 A, B). The data for the remaining groups is in accordance with a reduction in heterozygosity with aneuploidy turnover. However, to establish presence and effect sizes of a reduction in heterozygosity due to aneuploidy turnover direct experiments and more accurate estimates of aneuploidy turnover are needed, particularly using *in vivo* parasites.

Genomic signatures of hybridisation

To clarify the relationship between the high heterozygosity of some isolates, their phylogenetic position and the signatures of admixture, we examined the genomes of all 46 isolates with genome-wide heterozygosity greater than 0.004 in more detail for signs of past hybridisation (Fig. 3 A, row A1 in Table 1). This threshold was chosen to include the majority of samples that had putative hybrid ancestry in the admixture analysis, including the Çukurova samples of known hybrid origin (Rogers et al., 2014). The few isolates with lower heterozygosity but other evidence of admixture were also investigated (BPK512A1, L60b, CL-SL and OVN3 between groups, and LRC-L1311, LRC-L1312 and LRC-L1313 between subgroups; rows A2 & B6 in Table 1), but identifying details beyond admixture results was difficult with only little SNP information available (e.g. Fig. S5 A, S6 D). For the 46 high-heterozygosity isolates (Table 1), we inspected the distribution of heterozygous sites along each genome, looked for blocks of co-inherited variants and investigated patterns of allele-specific read coverage (i.e. sample allele frequency) across each chromosome. We also inferred maxicircle kinetoplast (mitochondrial) genome sequences: as kDNA is considered to be uniparentally inherited (Akopyants et al., 2009; Inbar et al., 2013), the phylogeny for these sequences should identify one parent of any hybrid isolates.

28 of the 46 high heterozygosity isolates appeared to represent genuine hybrid lineages (rows B1, B2 and B4 in Table 1), and for 17 of these, likely parents could be assigned (row B2 in Table 1). The largest group with identified parents is the Turkish isolates from Çukurova province (Rogers et al., 2014). Additionally, two Cypriot isolates (CH32 and CH34) showed patches of homozygosity closely related to the remaint Cypriot isolates and the Turkish CUK hybrids (Fig. 4, S5). Therefore, CH32 and CH34 likely represent hybrids closely related to the CUK hybrids, but clearly derived from an independent hybridisation event to the CUK population itself (Fig. 1 A). Another Turkish isolate (EP) appeared to have a similar evolutionary history with putative parental strains from the Linf1 and the CUK hybrids

(Fig. 4). In contrast to previous hybrids, for EP, there were entire homozygous chromosomes that resembled either of the two putative parental groups (chromosomes 4, 12, 22 and 32 for one and 11, 23 and 24 for the other parent; Fig. 4). Phylogenetic analysis of the kDNA maxicircles further showed identical sequences to the Cypriot hybrid samples (CH23 and CH34, Fig. S7, Table S2). Additionally, on two chromosomes, 5 and 31, allele frequency distributions in this isolate were not compatible with a single, clonal population of cells suggesting the presence of a second but very closely related low frequency clone in this sample (Fig. S6, S8). We also saw discrete patches of heterozygous and homozygous variants in two isolates from East Africa (GE and LEM3472) and one from Israel (LRC-L740) that did not fit into any of the main *L. donovan*i groups. These isolates appeared admixed between the North Ethiopia / Sudan group (Ldon3) and the *L. donovani* group present in the Middle East (Ldon4) (Fig. 1 A, 4, S5). For sample GE, kDNA further confirmed that one putative parent came from the Ldon3 group (Fig. S7). All the isolates from the Ldon3 group, were also highly heterozygous and so potentially hybrids, but we cannot exclude other possible origins for this heterozygosity (Fig. 4, S5, Table 1).

While the CUK samples are known to be of hybrid origin between a JPCM5-like *L. infantum* isolate and an unidentified parasite from the *L. donovani* complex (Rogers et al., 2014), our admixture results did not suggest hybridisation between genetic groups present in our dataset. This still held when varying *K* (the specified number of subpopulations) from 2 to 25 (Fig. S9). We therefore took a haplotype-based approach to increase the power to identify putative parents of these hybrids similar to that in Rogers *et al.* (2014), but now compared them to our larger set of isolates. We identified the largest homozygous regions in the CUK genomes: i.e. those that were either almost devoid of SNP differences to the JPCM5 reference genome or those that had a high density of fixed differences but lacked heterozygous sites, and generated phylogenies for these regions (see Material & Methods). Trees for the four largest regions (155 kb – 215 kb) placed the JPCM5-like parent close to *L. infantum* samples from China, rather than to the classical MON-1 and non-MON1 Mediterranean subgroups (Fig. S10 A, S1). Trees for the putative other parent always grouped CUK with CH samples similarly to the phylogeny of the maxicircle DNA (Fig. S7), suggesting these as closest putative parents to the CUK group in our sample collection (Fig. S10 B). The phylogenetic origin of the CH samples, however, still remained uncertain: in these four phylogenies the CH samples clustered twice next to the Ldon4 group, once next to Linf1 and once between both species. A haplotype-based approach as used for the CUK samples, and polarizing on several different isolates also did not give clear results (data not shown).

Isolates with genetically distinct (sub-)clones

Unexpectedly, for 12 of the remaining isolates (rows B3 – B5 in Table 1), many of the heterozygous sites were present at extreme (high / low) allele frequencies (11 isolates) or at multiple intermediate frequencies (isolate GILANI), incompatible with the allele frequencies estimated based on chromosomal somy (Fig. S6, S8). We suspect that these isolates represent a mixture of multiple cell clones. However, as low frequency variants are more at risk of being false positive SNP calls, we additionally selected a subset of the highest confidence SNPs to verify the observed frequency patterns (see Material & Methods). The MAM isolate had the highest heterozygosity in our collection: it only had 178 homozygous differences to the JPCM5 reference, but 50,534 heterozygous sites, with a frequency of the reference allele of ~0.92 across all chromosomes (Fig. S6 A). Phylogenies for inferred haplotypes of these low-frequency variants were closest but not part of the Ldon5 group (Fig. S11 A), although this was somewhat variable between chromosomes (Fig. S11 B-D). We concluded that the MAM sample is most likely a mixture between a JPCM5-like *L. infantum* strain at high (~0.92) and an *L. donovani* related to Ldon5 at low (0.08) sample frequency. Due to the low frequency of the 2nd strain it might be that alleles have been missed for SNP calling and therefore the calculated sample heterozygosity is lower than expected for interspecies F1 crosses (see Fig. S12). Similarly, the few heterozygous isolates within several *L. donovani* groups, BPK157A1 in Ldon1, Malta33 and GILANI in Ldon3, SUKKAR2 and BUMM3 in Ldon4 and LRC-L53 in Ldon5 (Fig. 3 A) all appeared to be mixtures of two clones from within the respective group (Fig. S11) apart from GILANI, which might be a more complex mixture (Fig. S6). For two of those samples the high number of within sample SNPs is due to segregating clones at high and low frequency (BPK157A1, LRC-L53 see row B3 in Table 1). For the other samples (BUMM3, Malta33, SUKKAR2; row B4 in Table 1) the majority of SNPs comes from heterozygous sites of a putative hybrid with a smaller fraction of SNPs owing to an additional related low frequency clone (Fig. S6). For isolate BPK157A1, the only sample in this subset re-grown from a single cell prior to sequencing (Table S1), is it contradictory to conclude that these variants are due to a mixture of clones. We ruled out false positive SNP calls by identifying 216 of the highest quality SNPs that show the extreme frequency pattern (Fig. S13; Material & Methods), however, alternate explanations including incomplete cloning or changes during *in vitro* culture post-cloning also seem unlikely. Highly heterozygous isolates from *L. infantum* (ISS174, ISS2426, ISS2429 and Inf152 in Linf1) also had skewed allele frequency distributions (Fig. S6), and therefore likely represent either mixed clone isolates or samples that have evolved significant diversity during *in vitro* growth. Samples, ISS174, ISS2426 and ISS2429, showed a strong positive correlation of chromosomal heterozygosity and somy not found in any other samples (Fig. S14). We speculate that these isolates may have

accumulated substantial numbers of new mutations most likely while maintaining relatively stable chromosome copy number during *in vitro* culture. Consequently, we expect relatively more mutations on chromosomes with a higher chromosome dosage, resulting in higher heterozygosity of high somy chromosomes.

Population genomic characterisation of the groups

Sexual recombination is not obligate in the *Leishmania* lifecycle and appears to be rare in many natural populations (Imamura et al., 2016; Ramírez and Llewellyn, 2014; Rougeron et al., 2009). We thus examined patterns of linkage disequilibrium (LD) between *Leishmania* populations as a clue to the frequency of sexual recombination, bearing in mind that LD can be affected by underlying population structure. LD estimates further depend on the frequency of recombination, the population size and the size of sample taken from the population (see also Fig. 5 A versus S15). We subsampled larger groups to identical groups sizes and found strong differences between groups in LD decay with genomic distance (Fig. 5 A). Linkage was strongest in the Ldon2 group with mean LD estimates around 0.9 regardless of genomic distance between SNPs, even when comparing sites on different chromosomes. The *L. infantum* groups (Linf1 and the CUK samples) started with high mean LD values for 1 kb distances above 0.9 and 0.8, respectively, and dropped down to ~0.5 for 100 kb distances and to ~0.4 and ~0.3 between chromosomes. Ldon3 and Ldon5 groups had the lowest LD estimates: at up to 1 kb distances LD had mean values of ~0.8 and 0.6 for Ldon3 and Ldon5, respectively, and dropped to ~0.2 for distances ≥50 kb in both groups and remained at those levels between chromosomes. All of these trends were relatively consistent among three independent subsamples from each of the larger groups, but the pattern was more complex for Ldon1. Here, the mean LD had a flat distribution with genomic distance like the Ldon2 group but at a much lower LD level, and showed significant variation between 3 subsamples (Fig. 5 B): two of the three subsamples showed low but very variable LD, and the third showed consistently high LD with distance. Low LD replicates were based on samples with a greatly reduced number of within-replicate SNPs (683 and 685 in R1 and R3 versus 23,303 SNPs in replicate R2). In the low LD replicates the majority of SNPs were singletons or present in only two copies, while in replicate R2 the majority of minor alleles were present at four copies (Fig. S16 A). Mean LD estimates across the entire Ldon1 group were also consistent at high levels above 0.8 independent of genomic distance (Fig. S15). We conclude that the substructure described for samples from the Indian subcontinent (Imamura et al., 2016) is responsible for varying LD estimates of the subsamples, with low LD replicates due to sampling only closely-related subgroups that only differ in

409 a small number of isolate-specific variants that are most parsimoniously described by recent
410 mutations (Fig. 5 B).
411
412 The groups also differed in their allele frequency distributions (i.e. the site frequency spectra, SFS). In
413 a diploid, panmictic and sexually recombining population of constant population size neutral sites
414 should segregate following a reciprocal function (Ferretti et al., 2018; Wright, 1938). While we would
415 not predict *Leishmania* populations to exactly follow these expectations, most of the groups (Ldon1,
416 Ldon2, Ldon5 and Linf1) were dominated by low frequency variants (Fig. S16). In contrast,
417 intermediate frequency variants were frequent in Ldon3 and even dominated variation in the *L.*
418 *infantum* CUK samples. The CUK group had been suggested to have largely expanded clonally from a
419 single hybridisation event between diverse strains with only small fractions of subsequent
420 hybridisation (Rogers et al., 2014). This scenario might explain why polymorphic sites, generated by
421 the hybridisation of diverse strains and common to the majority of samples can be at intermediate
422 population frequency. This group history also agrees with stronger LD for the short range due to
423 shared blocks that may be broken up by rare subsequent hybridisation and recombination events. For
424 the Ldon3 group increased intermediate frequency alleles combined with a strong decline of LD with
425 distance might suggest that old variants are segregating in the group at high frequencies, due to
426 relatively frequent hybridisation between clones within this group.
427
428 To identify genomic differences between the major groups, we determined the fixation index (F_{ST}) for
429 all SNP variants among pairs of groups, excluding samples identified as between group mixtures (Tab.
430 1 B3 & B4) or hybrids between groups (Tab. 1 B2, except CUK samples). Most SNP sites segregating
431 within each pair of groups were found to be population-specific, i.e. F_{ST}=1, in 10 out of 15 pairs (Fig. 6
432 A). This confirmed that most groups are well differentiated from each other with limited gene flow
433 between them. This high level of differentiation allowed us to identify between 6,769 and 26,145
434 potentially differentially fixed 'marker' SNPs for each group (Fig. 6 B, Tab. S3). These markers can be
435 useful in diagnosing parasite infections from particular groups, but might not be fixed in populations
436 identified based on a few isolates only. Despite this differentiation, many variants remained that were
437 fixed in combinations of groups. Most of these SNPs supported the species split, between *L. infantum*
438 and *L. donovani*, with 11,228 differentially fixed SNPs (Fig 6 C). Within-group genetic diversity varied
439 substantially between groups ranging from less than 1 SNP/10kb within the three CH samples to ~16
440 SNPs/10kb in Ldon4 (Fig. 6 D). Subsampled groups of seven isolates typically had ~3 SNPs/10kb, while
441 the two more polymorphic groups of *L. donovani* had SNP densities of ~12 and ~14 SNPs/10kb. Most
442 within-group segregating variation was group-specific: no SNPs segregated within all eight groups. The

443 most widespread polymorphisms are 4 SNPs shared between 6 groups and 25 SNPs segregating in at
444 least five of the eight groups and might be putative candidates for SNPs under balancing selection (Fig.
445 S17, Tab. S4).
446
447 **Copy number variation**
448
449 To assess the importance of genome structure variation in *Leishmania* evolution, we identified all large
450 sub-chromosome scale copy number variants (CNVs) within our isolates (duplications and deletions
451 ≥25 kb; see Material & Methods). In total, 940 large CNVs were found, an average of ~6 per sample.
452 75% of these large variants had a length ≤40 kb and only ~3% were >100 kb with the largest variant
453 of 675 kb (Table S5, Fig. S18). Most of these very large variants (>100 kb), were located on
454 chromosome 35 (Fig. S19). The frequency of large CNVs varied among chromosomes but was not
455 associated with chromosome length for duplications (Pearson correlation -0.06, p-value 0.74) and
456 showed a weak negative correlation for deletions (Pearson correlation 0.32, p-value 0.05) (Fig. S20).
457 We identified a total of 183 and 62 "unique" duplications and deletions, respectively, when clustering
458 each variant type across all samples based on chromosomal location (see Materials and Methods,
459 Table S6). Approximately half the CNVs were located at the chromosome ends, i.e. 22% and 26%
460 starting within 15 kb of chromosome 5' and 3' ends, respectively. The majority of large CNVs, were
461 present in only a single sample, but some were much more widespread – the most frequent being
462 present in 42 different samples and one variant being present in eight different groups (Fig. S21 A).
463 We were particularly interested in CNVs that were present in multiple groups or both species, as these
464 must either have been segregating over a long period of time, or have arisen multiple times
465 independently in different populations. 28% (69 of 245) of all variants were present in both species
466 (Fig. S21 B, Table S6) and we investigated those in more detail. We excluded terminal CNVs that
467 showed a gradual coverage increase towards the ends (e.g. Fig. S22) as these have been suspected to
468 be due to telomeric amplifications (Bussotti et al., 2018). Several other shared CNVs may represent
469 collapsed repeat regions in the reference genome assembly at which the repeat number varies
470 between samples or where coverage is close to our CNV coverage calling thresholds (e.g. Fig. S23), so
471 we inspected these manually. We describe in detail two examples of clear CNVs, one deletion and one
472 duplication. The 25 kb long deletion on chromosome 27 was present in 15% of all samples and across
473 four of the different identified groups including both species (Fig. 7 A). It always occurred on a disomic
474 background resulting in the loss of the allele. The 17 genes present within the deletion were enriched
475 for the GO term "cilium-dependent motility" (Fig. 7 C, Table S6). The duplication found on
476 chromosome 35 was only present in a single sample in each, the Ldon1 and Linf1, group (Fig. 7 B). In

Ldon1, it showed a 2-copy increase on a disomic background, suggesting it to be homozygous for a 1-copy duplication haplotype or heterozygous with one normal and one 2-copy duplication haplotype. In contrast, the sample from Linf1 has a single copy duplication on a trisomic background. 66 genes are present in the insertion enriched for several GO categories (Fig. 7 C).

To investigate smaller copy number variants, we determined the copy number (CN) for each gene in every sample by normalising the median gene coverage by the haploid coverage of the respective chromosome (see Material & Methods). CN variation affected 91.5% of genes (7,625 / 8330; Fig 9 A, Table S7), but most CNVs are rare (Fig. 8 A). Only 3.6% of all genes (304) showed a median copy number change ($\leq$-1 or $\geq$1) across samples with 103 genes decreased and 201 increased, respectively (Fig. 8 B). Enrichment tests for the 103 genes with frequently reduced copy number showed GO term enrichments for the biological processes “cation transport”, “transmembrane transport”, “fatty acid biosynthesis” and “localization” (median CN change across samples $\leq$1, Table S8). The 201 genes that were regularly increased showed enrichment for several terms including but not exclusive to “modulation by symbiont of host protein kinase-mediated signal transduction”, “cell adhesion” and “drug catabolic process” (median CN change across samples $\geq$1, for full list see Table S8). Only a subset of 52 genes (0.6%) showed frequently high gene copy number increases (median $\geq$4 across all samples). Enriched GO terms largely overlapped with enrichments of genes including small CN increases with the additional enrichment of “response to active oxygen species” (Table S8). Those categories might indicate functions on which there is frequent or strong selection pressure. Median gene copy number was positively correlated among groups (Fig. 8 C, Pearson correlation for pairwise comparisons between 0.8 and 0.91). Despite this extensive variation and shared copy number variation across groups, gene copy number still retained some phylogenetic signal (Fig. 8 D).

Genetic variation for known drug resistance loci

We investigated how genetic variation previously associated with drug resistance is distributed across our global collection of isolates, including loci involved in resistance to or treatment failure of antimonial drugs and Miltefosine (Table 2).

The best-known genetic variant associated with drug resistance in *Leishmania* is the so-called H-locus: amplification of this locus is involved in resistance to several unrelated drugs including antimonials (Callahan and Beverley, 1991; Dias et al., 2007; Grondin et al., 1993; Leprohon et al., 2009; Marchini et al., 2003). In our dataset, the four genes at this locus had an increased gene copy number in 30% of

511 the samples (CN +1 to +44) and a reduced copy number in 9% (CN -1; Table 2). 36% of all isolates had
512 a copy number increase of varying degree with identical insertion boundaries that included the genes
513 YIP1, MRPA and argininosuccinate synthase (Fig. 9 A, S23 A, Table 2, S7). This duplication was only
514 present in groups Ldon1 and Ldon3 with median increases of approximately +4 and +2, respectively.
515 This matches the rationale that parasites on the Indian subcontinent (largely Ldon1) have experienced
516 the highest drug pressure of antimonials in the past and are suggested to be preadapted to this drug
517 (Dumetz et al., 2018) and therefore have the highest prevalence and extent of CN increase, followed
518 by isolates from Sudan and Ethiopia (largely Ldon3). Under this scenario, the Pteridine reductase 1
519 gene at the H-locus may not be relevant for the drug resistance as it does not show an increased gene
520 CN along with the other genes at that locus (Fig. 9 A). One other isolate, LRC-L51p (Ldon5, India, 1954),
521 had a much larger duplication in this region including the entire H-locus and spanning >45 kb with an
522 enormous increase of ~+44 suggesting an independent insertion or amplification mechanism (Fig. S24
523 A). Four additional isolates showed a copy number increase for only two of the genes at the locus,
524 with different boundaries but always including the MRPA gene (Fig. S24 B).
525
526 Differential expression of the Mitogen-activated protein kinase 1 (MAPK1) has previously been
527 associated with antimony resistance. However, while (Singh et al., 2010) suggested that
528 overexpression is associated with resistance, (Ashutosh et al., 2012) suggest the opposite effect
529 potentially implicating an impact of the genetic background. As expression in *Leishmania* is typically
530 tightly linked with gene copy number (Barja et al., 2017; Iantorno et al., 2017), we summarised MAPK1
531 CNVs in our dataset (Table 2). 45% of all isolates had an amplified copy number at this locus, including
532 all isolates of Ldon1 and Ldon3 with the highest copy number increase between 12 and 41 copies in
533 Ldon1. Only a single *L. infantum* isolate had a reduced copy number of one (Fig. S25 A, Table 2, S7).
534 Increased copy number of MAPK1 is thus associated with isolates from geographical locations with
535 high historical antimonial drug pressures such as the Indian subcontinent and to a lesser extend Africa.
536 Additionally two isolates, BPK164A1 and BPK649A1 in Ldon1 belong to a highly Sb^V-resistant clade on
537 the Indian subcontinent (ISC5; Imamura *et al.*, 2016), which might argue that gene copy number
538 reduction is not the primary cause of expression decrease at this locus (e.g. Marquis *et al.*, 2005).
539 Another protein, the membrane channel protein aquaglyceroporin (AQP1), is known to be involved in
540 the uptake of pentavalent antimonials: reduced copy number and expression have been associated
541 with drug resistance (Andrade et al., 2016; Gourbal et al., 2004; Monte-Neto et al., 2015; Mukherjee
542 et al., 2013), as has other genetic variation at this locus (Imamura et al., 2016; Monte-Neto et al.,
543 2015; Uzcategui et al., 2008). In our dataset, copy number at this locus was reduced in 6% and
544 increased in 35% of all isolates with small effect sizes (CN -2 to -1 and +1 to +3) but at least one copy

of the locus was always present (Fig. S25 B, Table 2, S7). This may reflect resistance levels in the different populations, although direct sequencing from patient tissues has shown that structural variants might be lost during parasite isolation or subsequent growth *in vitro* potentially influencing our observations (Domagalska et al., 2019).

The Miltefosine transporter in *L. donovani* (LdMT) together with its putative ß subunit LdRos3 have been shown to be essential for phospholipid translocation activity and thereby the potency of the anti-leishmanial drug Miltefosine (Pérez-Victoria et al., 2006). In a drug selection experiment, Miltefosine resistant parasites showed common and strain-specific genetic changes including deletions at LdMT and single base mutations (Shaw et al., 2016). Neither LdMT, Ros3 or a hypothetical protein deleted together with LdMT in a drug selection experiment (Shaw et al., 2016), showed a reduction in gene copy number across our 151 isolates (Fig. S25 C, Table S7). Moreover, no SNP variation was present in two codons (A691, E197; Shaw *et al.*, 2016) putatively associated with drug resistance (Table 2). The Miltefosine sensitivity locus (MSL) was recently identified as a deletion associated with treatment failure in a clinical study of patients with VL in Brazil (Carnielli et al., 2018). In the same study, further genotyping of the MSL showed clinal variation in the presence of the locus ranging from 95% in North East Brazil to <5% in the South East (N=157), while no deletion was found in the Old World. The entire locus including all four genes (Table 2) was completely deleted in four of our samples of the Linf1 group including two of the four samples from Brazil (Cha001 1974, WC 2007) and in the two samples from Honduras (HN167 1998, HN336 1993) (Fig. 9 B, Table S7) with deletion boundaries coinciding with those reported previously (Carnielli et al., 2018). Another isolate, IMT373cl1 (Portugal, 2005) showed a deletion of a larger region (90 kb), reducing the local chromosome copy number from four to two (Fig. 9 B). The sixth sample that showed a copy number decrease of all four MSL associated genes, only showed a marginal and variable reduction in coverage and might be better explained by noise in genome coverage (Fig. 9 B).

Population and species-specific selection

We investigated putative species-specific selection, summarizing selection across the genome using the numbers of fixed vs. polymorphic and synonymous vs. non-synonymous sites for each species across all genes: The α statistic, originally by (Smith and Eyre-Walker, 2002), is a summary statistic, presenting the proportion of non-synonymous substitutions fixed by positive selection and is often used to summarize patterns of selection in a species. In both, *L. donovani* and *L. infantum*, α was negative, with -0.19 and -0.34, respectively, showing an excess of non-synonymous polymorphisms

579 but lacking a clear biological interpretation. Out of 8,234 genes tested for departure of neutrality using
580 the McDonald-Kreitman test, only two and four genes showed signs of positive selection (p-value
581 <0.05, FDR=1) and 11 and 12 an excess of non-synonymous differences (p-value <0.05, FDR=1) for *L.*
582 *donovani* and *L. infantum*, respectively (Fig. S26, Table S10). Interestingly, one of the genes with
583 putative signs of adaptive evolution in *L. donovani* (LINF_330040400 v41, LinJ.33.3220 v38) was
584 previously associated with *in vivo* enhanced virulence and increased parasite burden *in vitro* for *L.*
585 *major* when overexpressed (Reiling et al., 2010). In our dataset, this gene contained 9 missense, 3
586 synonymous and 19 upstream / intergenic SNP-variants differentially fixed between *L. donovani* and
587 *L. infantum* (Table S3), which might provide further candidates for differences in virulence between
588 both species.
589
590 While genetic variants can become fixed in different populations by either neutral forces (genetic drift)
591 or positive selection, we took advantage of the genetic differentiation between groups to search for
592 group-specific SNPs that might be of biological relevance. We investigated whether particular
593 functional categories (biological processes in Gene Ontology) were enriched among genes containing
594 high or moderate effect group- and species-specific SNP variants. While most enrichment terms were
595 specific to one marker set, the terms “protein phosphorylation”, “microtubule-based movement” and
596 “movement of cell or subcellular component” were enriched in five, three and two out of the nine
597 tested SNP sets, respectively (Fig. S27). More group specific enrichments with potentially more easily
598 interpretable biological implications include 1) “response to immune response of other organism
599 involved in symbiotic interaction” for Ldon1, 2) “mismatch repair” for Linf1 in response to oxidative
600 stress and 3) “pathogenesis” for the *L. infantum* – *L. donovani* species comparison (Fig. S27). For the
601 species comparison, the enrichment of the term “pathogenesis” was due to fixed differences of
602 putative functional relevance in genes including a protein containing a Tir chaperone (CesT) domain,
603 a subtilisin protease and a Bardet-biedl syndrome 1 protein that are putative candidates for increased
604 pathogenicity in *L. donovani* (Table 3, S3). Tir (translocated intimin receptor) chaperones are a family
605 of key indicators of pathogenic potential in gram-negative bacteria, where they support the type III
606 secretion system (Delahay et al., 2002). Proteins containing these domains are almost exclusive to
607 kinetoplastids among eukaryotes. In *L. donovani*, a subtilisin protease (SUB; Clan SB, family S8), has
608 been found to alter regulation of the trypanothione reductase system, which is required for reactive
609 oxygen detoxification in amastigotes and to be necessary for full virulence (Swenerton et al., 2010).
610 The Bardet-biedl syndrome 1 (BBS1) gene in *Leishmania* was shown to be involved in pathogen
611 infectivity. BBS1 knock-out strains, as promastigotes *in vitro*, had no apparent defects affecting
612 growth, flagellum assembly, motility or differentiation but showed a reduced infectivity for *in vitro*

macrophages and the ability to infect BALB/c mouse of null parasites was severely compromised (Price et al., 2013).

Discussion

Our whole-genome sequence data represents much of the global distribution of the *L. donovani* species complex. Compared to previous genomic studies on the *L. donovani* complex that focused on more geographically confined populations (Carnielli et al., 2018; Downing et al., 2011; Imamura et al., 2016; Rogers et al., 2014; Teixeira et al., 2017; Zackay et al., 2018), our sampling revealed a much greater genetic diversity. We identified five major clades of *L. donovani* that largely reflect the geographical distribution of the parasites and their associated vector species (Akhoundi et al., 2016). Some, such as the Middle Eastern group (Ldon4) are within themselves diverse, and in this case represented by a few samples, suggesting that a deeper sampling of parasites in this region may be needed. In contrast, our data confirmed that the low diversity of the main genotype group from the Indian subcontinent (Imamura et al., 2016) is indeed unusual, which might be related to the epidemic nature of VL on the Indian subcontinent (Dye and Wolpert, 1988). The main *L. infantum* clade is widespread and displays little diversity, although two subgroups represent the classical MON-1 and non-MON-1 Mediterranean lineages (Fig. 1 A, S1). Our data highlighted some weaknesses in previous typing systems for characterising *Leishmania* using MLEE (Rioux et al., 1990) and MLMT (Schönian et al., 2011, 2008). We confirmed paraphyly of the zymodeme MON-37 across *L. donovani* groups (see also Alam *et al.*, 2009) and for the zymodemes MON-30 and MON-82 within the Ldon3 group (Fig. S1). Moreover, the MON-1 zymodeme groups together parasites from the Mediterranean region and South America but also a sample from the genetically distinct Asian subgroup (Fig. S1). While data from MLMT (e.g. Kuhls *et al.*, (2007) and Gouzelou *et al.*, (2012) is much more congruent with our results, we explain diversity within the previously assigned Cypriot population (Gouzelou *et al.*, 2012) by hybridisation of some of these isolates (Fig. 1 A, 4, S5 A) and also describe hybridisation in other groups (e.g. LEM3472, GE and LRC-L740) that was not apparent with microsatellite markers (Kuhls et al., 2007).

Two regions emerged as apparent hot-spots of diversity in this species complex. The first is the Eastern Mediterranean, where the high genetic diversity of parasites assigned to *L. infantum* appears to be driven by hybridisation between *L. infantum* from China and a genotype identified in Cyprus (CH33,

35 and 36) (Fig. S10). This gave rise to the type of isolates from Çukurova described previously (Rogers et al., 2014) and some other hybrid genotypes from Cyprus (CH32 and 34). The phylogenetic origin of the five Cypriot isolates has been unclear: they were placed in the paraphyletic zymodeme MON-37 of *L. donovani* (Antoniou et al., 2008) but clustering based on microsatellite profiles placed them in a clade of *L. infantum* between zymodeme MON-1 and non-MON-1 isolates (Gouzelou et al., 2012). Our data supports a deep-branching clade of CH and CUK isolates distinct from other isolates of *L. infantum* (Fig, 1 A, S1) but the precise phylogenetic position of this group varies somewhat for different parts of the genome (Fig. S10 B). The origin of the pure, i.e. "non-hybrid" Cypriot samples (CH33, 35, 36), however, is not completely resolved: they could be either a distinct evolutionary linage within the *L. donovani* complex, or ancient hybrids between *L. infantum* and *L. donovani*. The other geographical regions of high diversity within the *L. donovani* complex is further South, encompassing the horn of Africa, the Arabian Peninsula and adjacent areas of the Middle East. Some of this diversity has been reported showing the presence of two clearly distinct groups of *L. donovani*: one in North-East and the other one in East Africa (Zackay et al., 2018). This genetic differentiation between both populations corresponds to their geographic separation by the rift valley in Ethiopia with different ecology and vector species (Gebre-Michael et al., 2010; Gebre-Michael and Lane, 1996) but hybrids between these populations have also been described (Cotton et al., 2019). More striking is the high diversity of *L. donovani* lineages in the Arabian Peninsula and the Middle East, including lineages present on both sides of the Red Sea and hybrids between groups present in this region and Africa (Ldon4 and other Ldon). The Middle East and adjacent regions may represent a contact zone where European, African and Asian lineages meet and occasionally hybridise increasing local genetic diversity. More extensive sampling in both of these 'hot-spot' regions would likely further improve our knowledge of the genetic diversity within the *L. donovani* species complex. Besides these 'diversity hot-spots', many other regions were sparsely sampled for our data collection and are under-explored by *Leishmania* researchers in general. We have few isolates from the New World, where VL is present in much of Central America, and northern South America (but see Teixeira *et al.*, 2017; Carnielli *et al.*, 2018), and almost none from Central Asia, where both *L. infantum* and *L. donovani* may be present. From China we only have *L. infantum* isolates, but there is likely to be a diverse range of *L. donovani*-complex parasites present (Alam et al., 2014; Zhang et al., 2013).

While we identified many novel lineages that are hybrids between major groups present in our study, it is likely that even with whole-genome variation data we are missing other admixture events especially within groups: This is because admixture analysis is most suited to identify admixed samples between the given *K* groups, and heterozygosities are most prominent when hybridisation occurs

between genetically diverse strains. All of our known hybrid populations had elevated levels of heterozygosity, but group Ldon3 was highly heterozygous without distinct genomic patterns of hybridisation (Fig. 3 A), so the generality of the relationship between heterozygosity and hybrid origin remains unclear. We investigated evidence from the admixture analysis (Fig. 1 A) at a range of values of the parameter *K* (the number of distinct populations present in the data; Fig. S9), also considering that many of the assumptions of admixture analysis are likely not to hold in *Leishmania* populations. However, this approach missed the known hybrids of the Çukurova population, which were consistently identified as a separate, "pure" population (Fig. S9). Therefore, we used an approach similar to that used by Rogers *et al.* (2014) to identify genome regions that seem to be homozygous for each of the two putative parents of hybrid groups. While this haplotype-based approach could identify parents of the Çukurova isolates, it did not clearly resolve the origins of other samples suggested to be hybrid by the admixture analysis. This could be either because our sample collection does not include the parental lineage or a close relative, or because these samples are of much older hybrid origin, so that subsequent recombination has erased the haplotype block structure we are looking for (e.g. see Rogers *et al.*, 2014). Different approaches are therefore needed to investigate recombination within populations. We also used the level of linkage disequilibrium as a measure for the level of recombination to show that the impact of recombination differs greatly between *L. donovani* complex populations. Additionally, we observed major differences in the allele frequency spectrum in different populations, in agreement with putative recombination differences and the unique evolutionary history of each group.

The variation in coverage between chromosomes and unusual allele frequency distributions in our isolates (Fig. S6, S8) confirmed the presence of extensive aneuploidy in our samples, as observed for all *Leishmania* promastigote cultures investigated to date. In our study, this variation in aneuploidy between samples reflected differences in the average chromosome copy number of a population of promastigote cells grown *in vitro* for each isolate, and showed no apparent phylogenetic structure. We assume that this reflects the well-documented mosaic aneuploidy present across *Leishmania* populations (Barja et al., 2017; Lachaud et al., 2014; Sterkers et al., 2011), where aneuploidy variation is present between cells within a parasite population. This variation could be selected upon and quickly change mean observed aneuploidies in a new environment, such as *in vitro* culture. However, we cannot directly address aneuploidy mosaicism with our data due to pooling cells within a strain for sequencing. To address this issue in future studies and understand the dynamics of *Leishmania* aneuploidy in infections and in culture, single-cell approaches seem to be most promising (e.g. Dujardin *et al.*, 2014).

Similarly, our data reflects the genetic variability of a set of isolates grown as promastigotes in axenic culture *in vitro*, a very different environment, and different life stage of the parasite to that present in patients. This means that we may miss variation present within host parasite populations that are lost during parasite isolation or subsequent growth, and that our results may be affected by selection to *in vitro* environments: In particular aneuploidy patterns in vectors and mammalian hosts were shown to differ from that in culture (Domagalska et al., 2019; Dumetz et al., 2017), and have other variants in particular during long term *in vitro* adaptation (e.g. Sinha *et al.*, 2018; Bussotti *et al.*, 2018). Given the breadth of global isolate collection used in our study it was not possible for us to ensure that common culture conditions were used for all the isolates. A recent approach to directly sequence *Leishmania* genomes in clinical samples has given some first insights into the effects of parasite culture *in vitro* and will allow future studies of *Leishmania* genome variation to avoid this potential bias (Domagalska et al., 2019).

Changes in gene dosage – of which aneuploidy is just the most striking example – have been shown to have a profound impact on gene expression in *Leishmania*, which lacks control of transcription initiation (Campbell et al., 2003). We identified extensive copy number variation, including both very large structural duplications and deletions and smaller-scale variants affecting single genes. Large structural variants are particularly common on chromosome 35. Many CNVs appeared too widespread across different clades to have evolved neutrally. While it is difficult to identify the specific functional relevance of these variants without phenotypic or functional information, these might be interesting targets for future functional studies. Additionally, we demonstrated the utility of genome data to understand functional genetic variation for variants with previously known impacts on phenotypes such as drug resistance. The deletion at the MSL locus, previously associated with Miltefosine treatment failure, is restricted to the New World and was considered to have evolved within Brazil (see also Carnielli *et al.*, 2018) but for the first time we reported this variant in Honduras, suggesting a geographically wider distribution than previously appreciated. Moreover, varying local frequencies and copy numbers of the H-locus and the MAPK1 duplication in India and North East Africa suggest that resistance against antimonials is more widespread on the Indian subcontinent, and may mediate a higher level of resistance than in other locations.

Our study provides the first comprehensive view of the globally distributed, whole-genome genetic diversity of the two most pathogenic species of *Leishmania* and any *Leishmania* species to date, and

748 will be a valuable resource in investigating individual loci to understand functional variation as well as
749 placing more focused studies into a global context.
750
751
752
753
754 **Material & Methods**
755
756 Choice of samples & sample origin
757 The genetic diversity of 151, mostly clinical isolates, from the *L. donovani* complex, and spanning the
758 entire global distribution of this species complex was investigated to reveal the complex's whole-
759 genome diversity on a global scale. This includes 98 isolates that we sequenced specifically for this
760 study, complemented with whole-genome sequence data of 33 isolates from the Indian subcontinent
761 (Imamura et al., 2016), 11 from a known Turkish hybrid population (Rogers et al., 2014), 6 from
762 Ethiopia (Zackay et al., 2018), 2 from Sri Lanka (Zhang et al., 2014) and the whole-genome sequences
763 of the JPCM5 reference strain (Peacock et al., 2007). All metadata on the 151 isolates used in this
764 study are summarized in Table S1 (see also https://microreact.org/project/_FWlYSTGf; Argimón et al.,
765 2016). The promastigote cultures and DNA samples came from different *Leishmania* strain collections:
766 The London School of Hygiene and Tropical Medicine; The Hebrew University, Jerusalem WHO
767 Reference Centre for the Leishmaniases; The Academic Medical Centre (University of Amsterdam),
768 Medical Microbiology, Section Parasitology; The Bangladesh Agricultural University, Mymensingh; The
769 Centre National de Référence des Leishmanioses Montpellier; The Istituto Superiore die Sanità Roma;
770 The Hellenic Pasteur Institute Athens; The Koret School of Veterinary Medicine, Hebrew University,
771 Jerusalem, Israel; The Coleção de *Leishmania* do Instituto Oswaldo Cruz, Rio de Janeiro; The University
772 of Khartoum; The Universitat Autònoma de Barcelona; The Institute of Tropical Medicine Antwerp,
773 and The Charité University Medicine Berlin. Only previously collected isolates from humans and
774 animals have been used in this study. The parasites from human cases had been isolated as part of
775 normal diagnosis and treatment with no unnecessary invasive procedures and data on human isolates
776 were encoded to maintain anonymity.
777
778 Whole-genome sequencing of clinical isolates
779 The 98 isolates new to this study were grown as *in vitro* promastigote culture to generate material for
780 sequencing as had been done for the 53 remaining sequenced isolates taken from other sources
781 (Imamura et al., 2016; Peacock et al., 2007; Rogers et al., 2014; Zackay et al., 2018; Zhang et al., 2014).

Of all these, most (62%) were not cloned and regrown from a single cell before sequencing; 6% of the isolates had been cloned and 32% were of unknown status prior to sequencing (Table S1). Genomic DNA was extracted by the phenol-chloroform method and quantified on a Qubit (Qubit Fluorometric Quantitation, Invitrogen, Life Technologies). DNA was then sheared into 400–600-base pair fragments by focused ultrasonication (Covaris Adaptive Focused Acoustics technology, AFA Inc., Woburn, USA). Standard indexed Illumina libraries were prepared using the NEBNext DNA Library Prep kit (New England BioLabs), followed by amplification using KAPA HiFI DNA polymerase (KAPA Biosystems). 100 bp paired-end reads were generated on the Illumina HiSeq 2000 according to the manufacturer's standard sequencing protocol (Bronner et al., 2014).

Read mapping pipeline

Reads were mapped with SMALT (v0.7.4, Ponstingl, 2010) using the parameters: "–x –y 0.9 –r 1 –i 1500" specifying independence of paired-end reads, a minimum fraction of 0.9 of matching bases, reporting of a random best alignment if multiple are present and a maximum insert size of 1500 bp against the reference genome JPCM5 of *L. infantum* (MCAN/ES/98/LLM-877, version v38 http://tritrypdb.org/tritrypdb/, Aslett et al., 2010). Mapped reads were sorted and duplicate reads were marked with picard "MarkDuplicates" (v1.92, https://broadinstitute.github.io/picard/). For resulting individual bam files per isolate, indels were called and local realignment was performed with GATK using the "RealignerTargetCreator" and "IndelRealigner" with default settings (v2.6-4, DePristo *et al.*, 2011).

Reference Genome Masking

We developed a custom mask for low complexity regions and gaps in the reference genome. To identify low complexity regions, we used the mappability tool from the GEM library (release3, Derrien et al., 2012). Gem-mappability was run with the parameters -l 100 -m 5 -e 0 --max-big-indel-length 0 --min-matched-bases 100, specifying a kmer length of 100 bp with up to 5 bp mismatches. This gives the number of distinct kmers in the genome, and we calculated the uniqueness of each bp position as the average number of kmers mapping a bp position. Any base with a GEM uniqueness score >1 was masked in the reference genome including a flanking region of 100 bp at either side. This approach masked 12.2 % of the 31.9 Mb genome.

Determination of sample ploidies

To determine individual chromosome ploidies per isolate the GATK tool "DepthOfCoverage" (v2.6-4) was used to obtain per-base read depth applying parameters: "--omitIntervalStatistics

--omitLocusTable --includeRefNSites --includeDeletions --printBaseCounts". Results files were masked using our custom mask (see "Reference Genome Mask"). Summary statistics were calculated per chromosome, including median read depth. The median read depth for each chromosome was used to estimate chromosome copy number, somy, for each sample using an Expectation-Maximization approach previously described in Iantorno *et al.* (2017). For a few isolates where the coverage model appeared to be overfitting (high deviance values), somy estimates were manually curated by examining both coverage and allele frequency data. Where allele frequency distributions did not support high somy values, they were altered so that the majority of chromosomes were disomic and individual errors were corrected to fit clear somy expectations suggested by the respective allele frequency spectra.

Variant calling

Variant calling was done following the Genome Analysis ToolKit (GATK) best-practice guidelines (Van der Auwera et al., 2013) with modifications detailed below. Given the aneuploidy of *Leishmania*, we considered individual somies per chromosome and isolate: the GATK "HaplotypeCaller" (v3.4-0) was used with the parameters "--sample_ploidy SOMY -dt NONE –annotateNDA" and additionally all-sites files were generated by adding the additional flag "-ERC BP_RESOLUTION" to the above HaplotypeCaller command. Individual vcf files (by chromosome and isolate) were processed, filtered and combined with custom made scripts implementing the following steps: only SNPs outside masked regions (see "Reference Genome Masking") were extracted; SNPs were hard filtered excluding genotypes failing to pass at least one of the following criteria: DP >= 5*SOMY, DP <= 1.75*(chromosome median read depth), FS <= 13.0 or missing, SOR <= 3.0 or missing, ReadPosRankSum <= 3.1 AND ReadPosRankSum >= -3.1, BaseQRankSum <= 3.1 AND BaseQRankSum >= -3.1, MQRankSum <= 3.1 AND MQRankSum >= -3.1, ClippingRankSum <= 3.1 AND ClippingRankSum >= -3.1. An additional masking was applied, based on the all-sites base quality information output by GATK HaplotypeCaller: DP >= 5*SOMY, DP <= 1.75*(chromosome median read depth) and GQ >= 10. Resulting samples were combined and SNPs with all reference or missing genotypes were removed.

Phylogenetic reconstruction

For phylogenetic reconstruction from whole-genome polymorphism data, all 395,602 SNPs that are polymorphic within the species complex and have a maximum fraction of 0.2 non-called sites across all 151 samples were considered. Nei's distances were calculated for bi-allelic sites per chromosome with the R package StAMPP (v1.5.1, Pembleton *et al.*, 2013), which takes into account aneuploidy across samples. Resulting distances matrices of Nei's distances per chromosome were weighted by

chromosomal SNP count forming a consensus distance matrix, that was used for phylogenetic reconstruction with the Neighbor-Joining algorithm implemented in the R package APE (v5.2, Saitou and Nei, 1987). For rooting of the tree, the phylogenetic reconstruction was repeated using three additional outgroup samples, of *L. major* (LmjFried, ENA: ERS001834), *L. tropica* (P283, ENA: ERS218438) and *L. mexicana* (LmexU1103 v1, ENA: ERS003040) (https://www.ebi.ac.uk/ena) using a total of 1,673,461 SNPs. Bootstrap values were calculated sampling 10 kb windows with replacement for a total of 1000 bootstrap replicates.

Phylogenetic reconstruction of maxicircles

Sequence reads were mapped against the maxicircle DNA of the reference strain, LV9 (MHOM/ET/1967/HU3), of *L. donovani* (TriTrypDB, with SMALT (v0.7.4, Ponstingl, 2010) using parameters: "-x -y 0.8 -r -1 -i 1500" and duplicates were marked with picard, "MarkDuplicates" (v1.92, https://broadinstitute.github.io/picard/). Local indel realignments were performed on the resulting alignments with GATK using the "RealignerTargetCreator" and "IndelRealigner" with default settings (v3.4-0, DePristo *et al.*, 2011) and subsequently filtered for a mapping quality of 20 and proper pairs using samtools, parameters "-q 20 -f 0x0002 -F 0x0004 -F 0x0008" (v1.3, Li *et al.*, 2009). SNP and Indel variants were called, hard filtered, selected and transformed to fasta sequences using GATK tools HaplotypeCaller, VariantFiltration, and FastaAlternateReferenceMaker (v3.4-0, DePristo et al., 2011). Used parameters include: "--sample_ploidy 1 -dt NONE –annotateNDA" (HaplotypeCaller), "QD < 2.0, MQ < 40.0, FS > 13.0, SOR > 4, BaseQRankSum > 3.1 || BaseQRankSum <-3.1", ClippingRankSum > 3.1 || ClippingRankSum <-3.1, MQRankSum > 3.1 || MQRankSum <-3.1, ReadPosRankSum > 3.1 || ReadPosRankSum <-3.1, DP > $DPmax, DP < $DPmin (SNP, VariantFiltration), "QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0" (Indel, VariantFiltration) and "-IUPAC 1" (FastaAlternateReferenceMaker). We determined maxicircle coverage of individual isolates using samtools depth (v1.3, Li *et al.*, 2009). Not all samples contained sufficient maxicircle DNA (likely depending on the DNA extraction protocol used) (Fig. 28). We therefore only used samples that had a medium coverage of at least 20, resulting in 116 samples (Fig. S7, Table S2) for subsequent analysis. As in the repetitive region of the maxicircle high quality mapping was not present, we assessed the minimum coverage across all 116 "good coverage" samples and based on that chose a region with a minimum coverage across those samples >=10 for subsequent alignment and phylogenetic reconstruction (positions 984 to 17162, Fig. S29). Resulting fasta sequences of individual maxicircles per isolates were aligned using MUSCLE (v3.8.31, Edgar, 2004) with default parameter settings and the phylogeny was reconstructed with RaxML (v7.0.3, Stamatakis, 2006) using parameters: "raxmlHPC -f a -m GTRGAMMA -p 12345 -x 12345 -# 100".

884

885 Gene-feature annotation and GO enrichment analysis

886 All SNPs were annotated with gene features using the software SNPeff (v4.2, Cingolani *et al.*, 2012).
887 Annotations for the reference genome *L. infantum*, JPCM5, were downloaded from TriTrypDB (v38,
888 http://tritrypdb.org/tritrypdb/; Aslett et al., 2010). Several gene sets of interest were subsequently
889 tested for Gene ontology (GO) term enrichments for the ontology "biological process". GO mappings
890 for *L. infantum* genes were downloaded from TriTryp DB (v38), where 4704 of the 8299 annotated
891 coding genes were also associated with a GO term. Enrichment of functional categories was tested
892 using the weightFisher algorithm in topGO (v2.34.0, Alexa et al., 2006) sing all genes annotated in the
893 "gene to GO" mapping file (v38). GO categories enriched with a p-value <0.05 (test: weightFish) were
894 subsequently visualised with Revigo (http://revigo.irb.hr/, assessed: February 2019, Supek et al.,
895 2011) using default settings and rectangle sizes normalized by absolute p-value.

896

897 Admixture analysis

898 To run ADMIXTURE (v1.23, Alexander et al., 2009), SNP genotype calls were collapsed from polysomic
899 to disomic for all chromosomes and only biallelic SNPs were included. SNPs were filtered and thinned,
900 removing SNPs with copies of the minor allele in less than 4 samples and one of two neighbouring
901 SNPs with a minimum distance < 250 bp. Using a five-fold cross-validation (CV) the optimal values of
902 *K* (smallest CV error) was determined to be 8 and 11 but we also explored different *K* values. The value
903 of *K* chosen was robust to different CV schemes.

904

905 Haplotype-based analysis of hybridisation in CUK isolates

906 We used SNP calls across all the original 12 CUK isolates from (Rogers et al., 2014) and called fractions
907 of heterozygous alleles and homozygous differences from the JPCM5 reference for 5 kb windows for
908 each isolate. Mean heterozygous and homozygous fractions per window were calculated as genomic
909 regions with either no SNP or increased number of homozygous differences (see also Rogers *et al.*,
910 2014). Putative parent blocks were identified using consecutive windows with mean heterozygous
911 fractions < 0.0002 (1 SNP/ 5 kb) and mean homozygous fractions either < 0.0004 (2 SNP/ 5 kb) for the
912 JPCM5-like parent or > 0.001 (5 SNP/ 5 kb) for the unknown parent. Those thresholds are quite
913 stringent (Fig. S30), but allowed conservative calling of putative parental haplotype regions. For each
914 parent, we selected the largest four regions conditioning on at most one block per chromosome
915 (resulting block sizes from 150 to 215 kb; Fig. S30). Phylogenetic trees for each of the eight regions
916 were then reconstructed based on polyploid genotypes of all 151 isolates and three outgroups
917 (LmjFried, *L. major*, ENA: ERS001834; P283, *L. tropica*, ENA: ERS218438; LmexU1103 v1, *L. mexicana*,

ENA: ERS003040; https://www.ebi.ac.uk/ena) using Nei's distances calculated with StAMPP (v1.5.1, Pembleton *et al.*, 2013) and the neighbour joining algorithm (R package ape, v5.2) in R (Supek et al., 2011).

Population genomics characterisation of the groups

For the population genomics characterization of the largest groups identified based on the global phylogeny (Fig. 1 A), isolates that were identified as putative mixtures of clones were removed. These were BPK157A1 (Ldon1), GILANI (Ldon3), LRC-L53 (Ldon5) and Inf152 (Linf1) and their respective groups are indicated by an asterisk (*). Polyploid genotype calls were transformed into diploid calls by transforming multiploid heterozygous sites into diploid heterozygous sites and polyploid homozygotes into diploid homozygotes. Linkage disequilibrium for each group was then calculated as genotype correlations of the transformed diploid calls for each of the identified groups using vcftools (v0.1.14, parameter: --geno-r2) (Danecek et al., 2011). F_{ST} between all group pairs was calculated for polymorphic sites with a minimum fraction of 0.8 called sites across all 151 samples as described in "Phylogenetic reconstruction" using the R package StAMPP (v1.5.1, Pembleton et al., 2013).

Genomic characterisation of individual isolates

Within isolate genome-wide heterozygosity was calculated using the formula: $1 - \frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{k_j} {p_{ij}}^2$ where p_i is the frequency of the i^{th} of k alleles for a given SNP genotype and the 1st summation sums over all *m* SNP loci for a given isolate. Here, genotype calls consider the correct somy for each isolate and chromosome as described above (see "Variant calling"). Isolate specific allele frequency spectra were obtained using mapped bam files including duplicate identification and indel realignment as described above (see "Read Mapping Pipeline"). Bam files were subsequently filtered using samtools view (v1.3, Li *et al.*, 2009) to only keep reads mapped in a proper pair with mapping quality of at least 20. Filtered bam files were summarised using samtools mpileup (v1.3, Li *et al.*, 2009) with arguments -d 3500 -B -Q 10 limiting the per sample coverage to 3500, disabling probabilistic realignment for the computation of base alignment quality and a minimum base quality of 10. The resulting mpileup file was converted to sync format summarising SNP allele counts per isolate using the mpileup2sync.jar script requiring a minimum base quality of 20 (Kofler et al., 2011). For the 11 samples with extreme allele frequency spectra, heterozygous SNPs were additionally filtered for the highest SNP calling quality of 99 (~10^{-10} probability of an incorrect genotype) and alternate alleles that were called as homozygous alternate alleles in at least five other isolates to confirm the presence of the skewed allele frequency spectra (Fig. S13).

952 Copy number variation

953 To identify large copy number variants (CNVs), realigned bam files for each sample were filtered for
954 proper-pairs and PCR or optical duplicates were removed using samtools view (v1.3, Li *et al.*, 2009).
955 Coverage was then determined using bedtools genomecov (v2.17.0) with parameters: "-d -split"
956 (Quinlan and Hall, 2010). Large duplications and deletion were identified using custom scripts in R (R
957 Core Team, 2013): genome coverage was determined for 5 kb non-overlapping windows along the
958 genome and each window was normalized by the haploid chromosome coverage of the respective
959 chromosome and sample (i.e. median chromosome coverage divided by somy of the respective
960 chromosome and sample). Large CNVs were identified through stretches of consecutive windows with
961 a somy-normalized median coverage >=0.5 or <=-0.5 for duplications and deletions, respectively, a
962 minimum length of 25 kb and a median normalized coverage difference across windows >=0.9 (Table
963 S5). To identify large CNVs across samples at identical positions and variant type, we grouped CNVs
964 across samples with identical start and end positions within <= 10 kb (i.e. up to two 5 kb windows
965 difference) (Table S6). CNVs of individual genes were determined based on the filtered bam files (see
966 genome coverages) with bedtools coverage (v2.17.0) using parameters "-d -split" (Quinlan and Hall,
967 2010) and analysing gene coverages in R (R Core Team, 2013). The coverage of each gene was
968 approximated by its median coverage and normalized by the haploid coverage of the respective
969 chromosome and sample (Table S7).
970
971 Measures of selection
972 For all genes with annotated mRNAs in TriTryp DB (v38, Aslett et al., 2010), the longest open reading
973 frames (ORF) were identified using a custom python script, resulting in 8,234 genes with and 5 without
974 ORFs. ORFs were then edited for SNP variation in both species using custom python scripts. Numbers
975 of polymorphic differences within a species versus fixed differences to an outgroup of both, non-
976 synonymous and synonymous sites, were annotated and tested for significance with Fisher's exact
977 test using previously implemented software (Holloway et al., 2007). This was done for each gene and
978 species always using the respective other species as an outgroup and removing sites polymorphic in
979 the outgroup. An unbiased version of the α statistic (Smith and Eyre-Walker, 2002; Stoletzki and Eyre-
980 Walker, 2011), intended to estimate the proportion of non-synonymous substitutions fixed by positive
981 selection across genes, was calculated with a custom R script.

Acknowledgements

We are very grateful to Gad **Baneth** from the Koret School of Veterinary Medicine, Hebrew University of Jerusalem, Israel; Patrick **Bastien** from the Centre National de Référence des Leishmanioses Montpellier, France; Sayda Hassan **El Safi** from Khartoum University, Khartoum, Sudan; Olga **Francino Martí** from Universitat Auònoma de Barcelona, Spain; Marina **Gramiccia** from the Infectious Diseases Department Istituto Superiore di Sanità, Rome, Italy; Ketty **Soteriadou** and Evi **Gouzelou** from the Hellenic Pasteur Institute Athens, Greece; A.K.M. **Shamsuzzaman** from Mymensingh Medical College and Be-Nazir **Ahmed** from Institute of Epidemiology, Disease Control and Research, Bangladesh; Vanessa **Yardley** from the London School of Hygiene & Tropical Medicine and Peter **Walden** from Charité Universitätsmedizin Berlin, Germany, each for providing isolates for this study. We also thank the *Leishmania* Collection of the **Oswaldo Cruz Foundation**, Brazil and Elisa **Cupolillo** for providing the isolates MHOM/BR/2003/MAM, MHOM/BR/2007/ARL, MHOM/BR/2007/WC, voucher numbers: IOC/L2651, IOC/L2935 and IOC/L3015, respectively. Further details of collaborators and their donated samples are given in Table S1.

This project was supported by Wellcome through its core funding of the Wellcome Sanger Institute (grants WT098051 and WT206194) and by the EU framework program FP7- 222895.

References

Akhoundi M, Kuhls K, Cannet A, Votýpka J, Marty P, Delaunay P, Sereno D. 2016. A Historical Overview of the Classification, Evolution, and Dispersion of *Leishmania* Parasites and Sandflies. *PLoS Negl Trop Dis* **10**:e0004349. doi:10.1371/journal.pntd.0004349

Akopyants NS, Kimblin N, Secundino N, Patrick R, Peters N, Lawyer P, Dobson DE, Beverley SM, Sacks DL. 2009. Demonstration of genetic exchange during cyclical development of *Leishmania* in the sand fly vector. *Science* **324**:265–268. doi:10.1126/science.1169464

Alam MZ, Haralambous C, Kuhls K, Gouzelou E, Sgouras D, Soteriadou K, Schnur L, Pratlong F, Schönian G. 2009. The paraphyletic composition of *Leishmania donovani* zymodeme MON-37 revealed by multilocus microsatellite typing. *Microbes Infect* **11**:707–715. doi:10.1016/j.micinf.2009.04.009

Alam MZ, Nakao R, Sakurai T, Kato H, Qu J-Q, Chai J-J, Chang KP, Schönian G, Katakura K. 2014. Genetic diversity of *Leishmania donovani/infantum* complex in China through microsatellite analysis. *Infect Genet Evol* **22**:112–119. doi:10.1016/j.meegid.2014.01.019

1019 Alemayehu B, Alemayehu M. 2017. Leishmaniasis: A Review on Parasite, Vector and
1020 Reservoir Host. *Health Sci J* **11**. doi:10.21767/1791-809X.1000519
1021 Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene
1022 expression data by decorrelating GO graph structure. *Bioinformatics* **22**:1600–1607.
1023 doi:10.1093/bioinformatics/btl140
1024 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in
1025 unrelated individuals. *Genome Res* **19**:1655–1664. doi:10.1101/gr.094052.109
1026 Alvar J, Vélez ID, Bern C, Herrero M, Desjeux P, Cano J, Jannin J, den Boer M. 2012.
1027 Leishmaniasis Worldwide and Global Estimates of Its Incidence. *PLoS ONE* **7**.
1028 doi:10.1371/journal.pone.0035671
1029 Andrade JM, Baba EH, Machado-de-Avila RA, Chavez-Olortegui C, Demicheli CP, Frézard F,
1030 Monte-Neto RL, Murta SMF. 2016. Silver and Nitrate Oppositely Modulate Antimony
1031 Susceptibility through Aquaglyceroporin 1 in *Leishmania* (Viannia) Species.
1032 *Antimicrob Agents Chemother* **60**:4482–4489. doi:10.1128/AAC.00768-16
1033 Antoniou M, Haralambous C, Mazeris A, Pratlong F, Dedet J-P, Soteriadou K. 2008.
1034 *Leishmania donovani* leishmaniasis in Cyprus. *Lancet Infect Dis* **8**:6–7.
1035 doi:10.1016/S1473-3099(07)70297-9
1036 Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG,
1037 Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and
1038 sharing data for genomic epidemiology and phylogeography. *Microb Genomics* **2**.
1039 doi:10.1099/mgen.0.000093
1040 Ashutosh, Garg M, Sundar S, Duncan R, Nakhasi HL, Goyal N. 2012. Downregulation of
1041 Mitogen-Activated Protein Kinase 1 of *Leishmania donovani* Field Isolates Is
1042 Associated with Antimony Resistance. *Antimicrob Agents Chemother* **56**:518–525.
1043 doi:10.1128/AAC.00736-11
1044 Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, Depledge DP,
1045 Fischer S, Gajria B, Gao X, Gardner MJ, Gingle A, Grant G, Harb OS, Heiges M, Hertz-
1046 Fowler C, Houston R, Innamorato F, Iodice J, Kissinger JC, Kraemer E, Li W, Logan FJ,
1047 Miller JA, Mitra S, Myler PJ, Nayak V, Pennington C, Phan I, Pinney DF, Ramasamy G,
1048 Rogers MB, Roos DS, Ross C, Sivam D, Smith DF, Srinivasamoorthy G, Stoeckert CJ,
1049 Subramanian S, Thibodeau R, Tivey A, Treatman C, Velarde G, Wang H. 2010.
1050 TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids*
1051 *Res* **38**:D457–D462. doi:10.1093/nar/gkp851
1052 Balloux F, Lehmann L, de Meeûs T. 2003. The population genetics of clonal and partially
1053 clonal diploids. *Genetics* **164**:1635–1644.
1054 Barja PP, Pescher P, Bussotti G, Dumetz F, Imamura H, Kedra D, Domagalska M, Chaumeau
1055 V, Himmelbauer H, Pages M, Sterkers Y, Dujardin J-C, Notredame C, Späth GF. 2017.
1056 Haplotype selection as an adaptive mechanism in the protozoan pathogen
1057 *Leishmania donovani*. *Nat Ecol Evol* **1**:1961. doi:10.1038/s41559-017-0361-x
1058 Bastien P, Blaineau C, Taminh M, Rioux JA, Roizès G, Pagès M. 1990. Interclonal variations in
1059 molecular karyotype in *Leishmania infantum* imply a 'mosaic' strain structure. *Mol*
1060 *Biochem Parasitol* **40**:53–61. doi:10.1016/0166-6851(90)90079-2
1061 Bhattarai NR, Van der Auwera G, Rijal S, Picado A, Speybroeck N, Khanal B, De Doncker S,
1062 Das ML, Ostyn B, Davies C, Coosemans M, Berkvens D, Boelaert M, Dujardin J-C.
1063 2010. Domestic Animals and Epidemiology of Visceral Leishmaniasis, Nepal. *Emerg*
1064 *Infect Dis* **16**:231–237. doi:10.3201/eid1602.090623

1065 Bronner IF, Quail MA, Turner DJ, Swerdlow H. 2014. Improved Protocols for Illumina
1066 Sequencing. *Curr Protoc Hum Genet* **80**:18.2.1-42.
1067 doi:10.1002/0471142905.hg1802s80
1068 Burza S, Croft SL, Boelaert M. 2018. Leishmaniasis. *The Lancet* **392**:951–970.
1069 doi:10.1016/S0140-6736(18)31204-2
1070 Bussotti G, Gouzelou E, Boité MC, Kherachi I, Harrat Z, Eddaikra N, Mottram JC, Antoniou M,
1071 Christodoulou V, Bali A, Guerfali FZ, Laouini D, Mukhtar M, Dumetz F, Dujardin J-C,
1072 Smirlis D, Lechat P, Pescher P, Hamouchi AE, Lemrani M, Chicharro C, Llanes-
1073 Acevedo IP, Botana L, Cruz I, Moreno J, Jeddi F, Aoun K, Bouratbine A, Cupolillo E,
1074 Späth GF. 2018. *Leishmania* Genome Dynamics during Environmental Adaptation
1075 Reveal Strain-Specific Differences in Gene Copy Number Variation, Karyotype
1076 Instability, and Telomeric Amplification. *mBio* **9**:e01399-18.
1077 doi:10.1128/mBio.01399-18
1078 Callahan HL, Beverley SM. 1991. Heavy metal resistance: a new role for P-glycoproteins in
1079 *Leishmania*. *J Biol Chem* **266**:18427–18430.
1080 Campbell DA, Thomas S, Sturm NR. 2003. Transcription in kinetoplastid protozoa: why be
1081 normal? *Microbes Infect* **5**:1231–1240. doi:10.1016/j.micinf.2003.09.005
1082 Carnielli JBT, Crouch K, Forrester S, Silva VC, Carvalho SFG, Damasceno JD, Brown E, Dickens
1083 NJ, Costa DL, Costa CHN, Dietze R, Jeffares DC, Mottram JC. 2018. A *Leishmania*
1084 *infanum* genetic marker associated with miltefosine treatment failure for visceral
1085 leishmaniasis. *EBioMedicine* **36**:83–91. doi:10.1016/j.ebiom.2018.09.029
1086 Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A
1087 program for annotating and predicting the effects of single nucleotide
1088 polymorphisms, SnpEff. *Fly (Austin)* **6**:80–92. doi:10.4161/fly.19695
1089 Cotton J, Durrant C, Franssen S, Gelanew T, Hailu A, Mateus D, Sanders M, Berriman M, Volf
1090 P, Miles M, Yeo M. 2019. Genomic analysis of natural intra-specific hybrids among
1091 Ethiopian isolates of *Leishmania donovani*. *bioRxiv* 516211. doi:10.1101/516211
1092 Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G,
1093 Marth GT, Sherry ST, McVean G, Durbin R. 2011. The variant call format and
1094 VCFtools. *Bioinformatics* **27**:2156–2158. doi:10.1093/bioinformatics/btr330
1095 De Meeûs T, Lehmann L, Balloux F. 2006. Molecular epidemiology of clonal diploids: A quick
1096 overview and a short DIY (do it yourself) notice. *Infect Genet Evol* **6**:163–170.
1097 doi:10.1016/j.meegid.2005.02.004
1098 Delahay RM, Shaw RK, Elliott SJ, Kaper JB, Knutton S, Frankel G. 2002. Functional analysis of
1099 the enteropathogenic Escherichia coli type III secretion system chaperone CesT
1100 identifies domains that mediate substrate interactions. *Mol Microbiol* **43**:61–73.
1101 doi:10.1046/j.1365-2958.2002.02740.x
1102 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Angel G
1103 del, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY,
1104 Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation
1105 discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*
1106 **43**:491. doi:10.1038/ng.806
1107 Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast
1108 Computation and Applications of Genome Mappability. *PLoS ONE* **7**.
1109 doi:10.1371/journal.pone.0030377
1110 Dias FC, Ruiz JC, Lopes WCZ, Squina FM, Renzi A, Cruz AK, Tosi LRO. 2007. Organization of H
1111 locus conserved repeats in *Leishmania (Viannia) braziliensis* correlates with lack of

gene amplification and drug resistance. *Parasitol Res* **101**:667–676. doi:10.1007/s00436-007-0528-5

Díaz-Sáez V, Merino-Espinosa G, Morales-Yuste M, Corpas-López V, Pratlong F, Morillas-Márquez F, Martín-Sánchez J. 2014. High rates of *Leishmania infantum* and *Trypanosoma nabiasi* infection in wild rabbits (*Oryctolagus cuniculus*) in sympatric and syntrophic conditions in an endemic canine leishmaniasis area: Epidemiological consequences. *Vet Parasitol* **202**:119–127. doi:10.1016/j.vetpar.2014.03.029

Domagalska MA, Imamura H, Sanders M, Broeck FV den, Bhattarai NR, Vanaerschot M, Maes I, D'Haenens E, Rai K, Rijal S, Berriman M, Cotton JA, Dujardin J-C. 2019. Genomes of intracellular Leishmania parasites directly sequenced from patients. *bioRxiv* 676163. doi:10.1101/676163

Downing T, Imamura H, Decuypere S, Clark TG, Coombs GH, Cotton JA, Hilley JD, de Doncker S, Maes I, Mottram JC, Quail MA, Rijal S, Sanders M, Schönian G, Stark O, Sundar S, Vanaerschot M, Hertz-Fowler C, Dujardin J-C, Berriman M. 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res* **21**:2143–2156. doi:10.1101/gr.123430.111

Dujardin J-C, Mannaert A, Durrant C, Cotton JA. 2014. Mosaic aneuploidy in *Leishmania*: the perspective of whole genome sequencing. *Trends Parasitol* **30**:554–555. doi:10.1016/j.pt.2014.09.004

Dumetz F, Cuypers B, Imamura H, Zander D, D'Haenens E, Maes I, Domagalska MA, Clos J, Dujardin J-C, De Muylder G. 2018. Molecular Preadaptation to Antimony Resistance in Leishmania donovani on the Indian Subcontinent. *mSphere* **3**. doi:10.1128/mSphere.00548-17

Dumetz F, Imamura H, Sanders M, Seblova V, Myskova J, Pescher P, Vanaerschot M, Meehan CJ, Cuypers B, De Muylder G, Späth GF, Bussotti G, Vermeesch JR, Berriman M, Cotton JA, Volf P, Dujardin JC, Domagalska MA. 2017. Modulation of Aneuploidy in *Leishmania* donovani during Adaptation to Different In Vitro and In Vivo Environments and Its Impact on Gene Expression. *mBio* **8**. doi:10.1128/mBio.00599-17

Dye C, Wolpert DM. 1988. Earthquakes, influenza and cycles of Indian kala-azar. *Trans R Soc Trop Med Hyg* **82**:843–850. doi:10.1016/0035-9203(88)90013-2

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797. doi:10.1093/nar/gkh340

Ferretti L, Klassmann A, Raineri E, Ramos-Onsins SE, Wiehe T, Achaz G. 2018. The neutral frequency spectrum of linked sites. *Theor Popul Biol* **123**:70–79. doi:10.1016/j.tpb.2018.06.001

Gebre-Michael T, Balkew M, Berhe N, Hailu A, Mekonnen Y. 2010. Further studies on the phlebotomine sandflies of the kala-azar endemic lowlands of Humera-Metema (north-west Ethiopia) with observations on their natural blood meal sources. *Parasit Vectors* **3**:6. doi:10.1186/1756-3305-3-6

Gebre-Michael T, Lane RP. 1996. The roles of *Phlebotomus martini* and *P. celiae* (Diptera: Phlebotominae) as vectors of visceral leishmaniasis in the Aba Roba focus, southern Ethiopia. *Med Vet Entomol* **10**:53–62. doi:10.1111/j.1365-2915.1996.tb00082.x

González-de la Fuente S, Camacho E, Peiró-Pastor R, Rastrojo A, Carrasco-Ramiro F, Aguado B, Requena JM, González-de la Fuente S, Camacho E, Peiró-Pastor R, Rastrojo A, Carrasco-Ramiro F, Aguado B, Requena JM. 2019. Complete and de novo assembly of

1159 the Leishmania braziliensis (M2904) genome. *Mem Inst Oswaldo Cruz* **114**.
1160 doi:10.1590/0074-02760180438
1161 Gourbal B, Sonuc N, Bhattacharjee H, Legare D, Sundar S, Ouellette M, Rosen BP,
1162 Mukhopadhyay R. 2004. Drug Uptake and Modulation of Drug Resistance in
1163 Leishmania by an Aquaglyceroporin. *J Biol Chem* **279**:31010–31017.
1164 doi:10.1074/jbc.M403959200
1165 Gouzelou E, Haralambous C, Amro A, Mentis A, Pratlong F, Dedet J-P, Votypka J, Volf P,
1166 Ozensoy Toz S, Kuhls K, Schönian G, Soteriadou K. 2012. Multilocus Microsatellite
1167 Typing (MLMT) of Strains from Turkey and Cyprus Reveals a Novel Monophyletic *L.*
1168 *donovani* Sensu Lato Group. *PLoS Negl Trop Dis* **6**. doi:10.1371/journal.pntd.0001507
1169 Grondin K, Papadopoulou B, Ouellette M. 1993. Homologous recombination between direct
1170 repeat sequences yields P-glycoprotein containing amplicons in arsenite resistant
1171 Leishmania. *Nucleic Acids Res* **21**:1895–1901.
1172 Guerbouj S, Guizani I, Speybroeck N, Le Ray D, Dujardin JC. 2001. Genomic polymorphism of
1173 *Leishmania infantum*: a relationship with clinical pleomorphism? *Infect Genet Evol*
1174 **1**:49–59. doi:10.1016/S1567-1348(01)00008-9
1175 Herrera G, Hernández C, Ayala MS, Flórez C, Teherán AA, Ramírez JD. 2017. Evaluation of a
1176 Multilocus Sequence Typing (MLST) scheme for *Leishmania (Viannia) braziliensis* and
1177 *Leishmania (Viannia) panamensis* in Colombia. *Parasit Vectors* **10**.
1178 doi:10.1186/s13071-017-2175-8
1179 Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. 2007. Adaptive Gene
1180 Expression Divergence Inferred from Population Genomics. *PLOS Genet* **3**:e187.
1181 doi:10.1371/journal.pgen.0030187
1182 Huang W, Massouras A, Inoue Y, Peiffer J, Rámia M, Tarone A, Turlapati L, Zichner T, Zhu D,
1183 Lyman R, Magwire M, Blankenburg K, Carbone MA, Chang K, Ellis L, Fernandez S, Han
1184 Y, Highnam G, Hjelmen C, Jack J, Javaid M, Jayaseelan J, Kalra D, Lee S, Lewis L,
1185 Munidasa M, Ongeri F, Patel S, Perales L, Perez A, Pu L, Rollmann S, Ruth R, Saada N,
1186 Warner C, Williams A, Wu Y-Q, Yamamoto A, Zhang Y, Zhu Y, Anholt R, Korbel J,
1187 Mittelman D, Muzny D, Gibbs R, Barbadilla A, Johnston S, Stone E, Richards S,
1188 Deplancke B, Mackay T. 2014. Natural variation in genome architecture among 205
1189 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* gr.171546.113.
1190 doi:10.1101/gr.171546.113
1191 Iantorno SA, Durrant C, Khan A, Sanders MJ, Beverley SM, Warren WC, Berriman M, Sacks
1192 DL, Cotton JA, Grigg ME. 2017. Gene Expression in *Leishmania* Is Regulated
1193 Predominantly by Gene Dosage. *mBio* **8**. doi:10.1128/mBio.01393-17
1194 Imamura H, Downing T, Van den Broeck F, Sanders MJ, Rijal S, Sundar S, Mannaert A,
1195 Vanaerschot M, Berg M, De Muylder G, Dumetz F, Cuypers B, Maes I, Domagalska M,
1196 Decuypere S, Rai K, Uranw S, Bhattarai NR, Khanal B, Prajapati VK, Sharma S, Stark O,
1197 Schönian G, De Koning HP, Settimo L, Vanhollebeke B, Roy S, Ostyn B, Boelaert M,
1198 Maes L, Berriman M, Dujardin J-C, Cotton JA. 2016. Evolutionary genomics of
1199 epidemic visceral leishmaniasis in the Indian subcontinent. *eLife* **5**:e12613.
1200 doi:10.7554/eLife.12613
1201 Inbar E, Akopyants NS, Charmoy M, Romano A, Lawyer P, Elnaiem D-EA, Kauffmann F,
1202 Barhoumi M, Grigg M, Owens K, Fay M, Dobson DE, Shaik J, Beverley SM, Sacks D.
1203 2013. The Mating Competence of Geographically Diverse *Leishmania major* Strains in
1204 Their Natural and Unnatural Sand Fly Vectors. *PLOS Genet* **9**:e1003672.
1205 doi:10.1371/journal.pgen.1003672

1206 Inbar E, Shaik J, Iantorno SA, Romano A, Nzelu CO, Owens K, Sanders MJ, Dobson D, Cotton
1207 JA, Grigg ME, Beverley SM, Sacks D. 2019. Whole genome sequencing of
1208 experimental hybrids supports meiosis-like sexual recombination in *Leishmania*.
1209 *PLOS Genet* **15**:e1008042. doi:10.1371/journal.pgen.1008042
1210 Jamjoom MB, Ashford RW, Bates PA, Chance ML, Kemp SJ, Watts PC, Noyes HA. 2004.
1211 *Leishmania donovani* is the only cause of visceral leishmaniasis in East Africa;
1212 previous descriptions of *L. infantum* and "*L. archibaldi*" from this region are a
1213 consequence of convergent evolution in the isoenzyme data. *Parasitology* **129**:399–
1214 409. doi:10.1017/S0031182004005955
1215 Kofler R, Pandey RV, Schlötterer C. 2011. PoPoolation2: identifying differentiation between
1216 populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics*
1217 **27**:3435–3436. doi:10.1093/bioinformatics/btr589
1218 Kuhls K, Keilonat L, Ochsenreither S, Schaar M, Schweynoch C, Presber W, Schönian G. 2007.
1219 Multilocus microsatellite typing (MLMT) reveals genetically isolated populations
1220 between and within the main endemic regions of visceral leishmaniasis. *Microbes*
1221 *Infect* **9**:334–343. doi:10.1016/j.micinf.2006.12.009
1222 Lachaud L, Bourgeois N, Kuk N, Morelle C, Crobu L, Merlin G, Bastien P, Pagès M, Sterkers Y.
1223 2014. Constitutive mosaic aneuploidy is a unique genetic feature widespread in the
1224 *Leishmania* genus. *Microbes Infect* **16**:61–66. doi:10.1016/j.micinf.2013.09.005
1225 Leblois R, Kuhls K, François O, Schönian G, Wirth T. 2011. Guns, germs and dogs: On the
1226 origin of *Leishmania chagasi*. *Infect Genet Evol* **11**:1091–1095.
1227 doi:10.1016/j.meegid.2011.04.004
1228 Leprohon P, Légaré D, Raymond F, Madore É, Hardiman G, Corbeil J, Ouellette M. 2009.
1229 Gene expression modulation is associated with gene amplification, supernumerary
1230 chromosomes and chromosome loss in antimony-resistant *Leishmania infantum*.
1231 *Nucleic Acids Res* **37**:1387–1399. doi:10.1093/nar/gkn1069
1232 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.
1233 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–
1234 2079. doi:10.1093/bioinformatics/btp352
1235 Lun Z-R, Wu M-S, Chen Y-F, Wang J-Y, Zhou X-N, Liao L-F, Chen J-P, Chow LMC, Chang KP.
1236 2015. Visceral Leishmaniasis in China: an Endemic Disease under Control. *Clin*
1237 *Microbiol Rev* **28**:987–1004. doi:10.1128/CMR.00080-14
1238 Lysenko AJa. 1971. Distribution of leishmaniasis in the Old World. *Bull World Health Organ*
1239 **44**:515–520.
1240 Marchini JFM, Cruz AK, Beverley SM, Tosi LRO. 2003. The H region HTBF gene mediates
1241 terbinafine resistance in *Leishmania major*. *Mol Biochem Parasitol* **131**:77–81.
1242 doi:10.1016/S0166-6851(03)00174-9
1243 Marquis N, Gourbal B, Rosen BP, Mukhopadhyay R, Ouellette M. 2005. Modulation in
1244 aquaglyceroporin AQP1 gene transcript levels in drug-resistant *Leishmania*. *Mol*
1245 *Microbiol* **57**:1690–1699. doi:10.1111/j.1365-2958.2005.04782.x
1246 McCall L-I, Zhang W-W, Matlashewski G. 2013. Determinants for the development of
1247 visceral leishmaniasis disease. *PLoS Pathog* **9**:e1003053.
1248 doi:10.1371/journal.ppat.1003053
1249 Monte-Neto R, Laffitte M-CN, Leprohon P, Reis P, Frézard F, Ouellette M. 2015.
1250 Intrachromosomal amplification, locus deletion and point mutation in the
1251 aquaglyceroporin AQP1 gene in antimony resistant *Leishmania (Viannia) guyanensis*.
1252 *PLoS Negl Trop Dis* **9**:e0003476. doi:10.1371/journal.pntd.0003476

1253 Mougneau E, Bihl F, Glaichenhaus N. 2011. Cell biology and immunology of *Leishmania*.
1254 *Immunol Rev* **240**:286–296. doi:10.1111/j.1600-065X.2010.00983.x
1255 Mukherjee A, Boisvert S, Monte-Neto RL do, Coelho AC, Raymond F, Mukhopadhyay R,
1256 Corbeil J, Ouellette M. 2013. Telomeric gene deletion and intrachromosomal
1257 amplification in antimony-resistant *Leishmania*. *Mol Microbiol* **88**:189–202.
1258 doi:10.1111/mmi.12178
1259 Ostyn B, Gidwani K, Khanal B, Picado A, Chappuis F, Singh SP, Rijal S, Sundar S, Boelaert M.
1260 2011. Incidence of Symptomatic and Asymptomatic Leishmania donovani Infections
1261 in High-Endemic Foci in India and Nepal: A Prospective Study. *PLoS Negl Trop Dis*
1262 **5**:e1284. doi:10.1371/journal.pntd.0001284
1263 Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, Quail MA, Peters N, Adlem E, Tivey A,
1264 Aslett M, Kerhornou A, Ivens A, Fraser A, Rajandream M-A, Carver T, Norbertczak H,
1265 Chillingworth T, Hance Z, Jagels K, Moule S, Ormond D, Rutter S, Squares R,
1266 Whitehead S, Rabbinowitsch E, Arrowsmith C, White B, Thurston S, Bringaud F,
1267 Baldauf SL, Faulconbridge A, Jeffares D, Depledge DP, Oyola SO, Hilley JD, Brito LO,
1268 Tosi LRO, Barrell B, Cruz AK, Mottram JC, Smith DF, Berriman M. 2007. Comparative
1269 genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat*
1270 *Genet* **39**:839–847. doi:10.1038/ng2053
1271 Pembleton LW, Cogan NOI, Forster JW. 2013. StAMPP: an R package for calculation of
1272 genetic differentiation and structure of mixed-ploidy level populations. *Mol Ecol*
1273 *Resour* **13**:946–952. doi:10.1111/1755-0998.12129
1274 Pérez-Victoria FJ, Sánchez-Cañete MP, Castanys S, Gamarro F. 2006. Phospholipid
1275 translocation and miltefosine potency require both L. donovani miltefosine
1276 transporter and the new protein LdRos3 in *Leishmania* parasites. *J Biol Chem*
1277 **281**:23766–23775. doi:10.1074/jbc.M605214200
1278 Ponstingl H. 2010. SMALT mapper, https://www.sanger.ac.uk/science/tools/smalt-0.
1279 *Httpswwwsangeracuksciencetoolssmalt-0*.
1280 https://www.sanger.ac.uk/science/tools/smalt-0
1281 Price HP, Paape D, Hodgkinson MR, Farrant K, Doehl J, Stark M, Smith DF. 2013. The
1282 *Leishmania major* BBSome subunit BBS1 is essential for parasite virulence in the
1283 mammalian host. *Mol Microbiol* **90**:597–611. doi:10.1111/mmi.12383
1284 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic
1285 features. *Bioinformatics* **26**:841–842. doi:10.1093/bioinformatics/btq033
1286 Quinnell RJ, Courtenay O. 2009. Transmission, reservoir hosts and control of zoonotic
1287 visceral leishmaniasis. *Parasitology* **136**:1915–1934.
1288 doi:10.1017/S0031182009991156
1289 R Core Team. 2013. R: A Language and Environment for Statistical Computing.
1290 Ramírez JD, Llewellyn MS. 2014. Reproductive clonality in protozoan pathogens—truth or
1291 artefact? *Mol Ecol* **23**:4195–4202. doi:10.1111/mec.12872
1292 Ready PD. 2014. Epidemiology of visceral leishmaniasis. *Clin Epidemiol* **6**:147–154.
1293 doi:10.2147/CLEP.S44267
1294 Real F, Vidal RO, Carazzolle MF, Mondego JMC, Costa GGL, Herai RH, Würtele M, de
1295 Carvalho LM, e Ferreira RC, Mortara RA, Barbiéri CL, Mieczkowski P, da Silveira JF,
1296 Briones MR da S, Pereira GAG, Bahia D. 2013. The Genome Sequence of *Leishmania*
1297 *(Leishmania) amazonensis*: Functional Annotation and Extended Analysis of Gene
1298 Models. *DNA Res Int J Rapid Publ Rep Genes Genomes* **20**:567–581.
1299 doi:10.1093/dnares/dst031

Reiling L, Chrobak M, Schmetz C, Clos J. 2010. Overexpression of a single *Leishmania major* gene enhances parasite infectivity in vivo and in vitro. *Mol Microbiol* **76**:1175–1190. doi:10.1111/j.1365-2958.2010.07130.x

Rioux JA, Lanotte G, Serres E, Pratlong F, Bastien P, Perieres J. 1990. Taxonomy of *Leishmania*. Use of isoenzymes. Suggestions for a new classification. *Ann Parasitol Hum Comp* **65**:111–125. doi:10.1051/parasite/1990653111

Rodrigues V, Cordeiro-da-Silva A, Laforge M, Silvestre R, Estaquier J. 2016. Regulation of immunity during visceral *Leishmania* infection. *Parasit Vectors* **9**. doi:10.1186/s13071-016-1412-x

Rogers MB, Downing T, Smith BA, Imamura H, Sanders M, Svobodova M, Volf P, Berriman M, Cotton JA, Smith DF. 2014. Genomic Confirmation of Hybridisation and Recent Inbreeding in a Vector-Isolated *Leishmania* Population. *PLOS Genet* **10**:e1004092. doi:10.1371/journal.pgen.1004092

Rogers MB, Hilley JD, Dickens NJ, Wilkes J, Bates PA, Depledge DP, Harris D, Her Y, Herzyk P, Imamura H, Otto TD, Sanders M, Seeger K, Dujardin J-C, Berriman M, Smith DF, Hertz-Fowler C, Mottram JC. 2011. Chromosome and gene copy number variation allow major structural change between species and strains of *Leishmania*. *Genome Res* **21**:2129–2142. doi:10.1101/gr.122945.111

Romano A, Inbar E, Debrabant A, Charmoy M, Lawyer P, Ribeiro-Gomes F, Barhoumi M, Grigg M, Shaik J, Dobson D, Beverley SM, Sacks DL. 2014. Cross-species genetic exchange between visceral and cutaneous strains of *Leishmania* in the sand fly vector. *Proc Natl Acad Sci U S A* **111**:16808–16813. doi:10.1073/pnas.1415109111

Rougeron V, Meeûs TD, Hide M, Waleckx E, Bermudez H, Arevalo J, Llanos-Cuentas A, Dujardin J-C, Doncker SD, Ray DL, Ayala FJ, Bañuls A-L. 2009. Extreme inbreeding in *Leishmania braziliensis*. *Proc Natl Acad Sci* **106**:10224–10229. doi:10.1073/pnas.0904420106

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**:406–425. doi:10.1093/oxfordjournals.molbev.a040454

Schönian G, Kuhls K, Mauricio IL. 2011. Molecular approaches for a better understanding of the epidemiology and population genetics of *Leishmania*. *Parasitology* **138**:405–425. doi:10.1017/S0031182010001538

Schönian G, Mauricio I, Gramiccia M, Cañavate C, Boelaert M, Dujardin J-C. 2008. Leishmaniases in the Mediterranean in the era of molecular epidemiology. *Trends Parasitol* **24**:135–142. doi:10.1016/j.pt.2007.12.006

Shaw CD, Lonchamp J, Downing T, Imamura H, Freeman TM, Cotton JA, Sanders M, Blackburn G, Dujardin JC, Rijal S, Khanal B, Illingworth CJR, Coombs GH, Carter KC. 2016. In vitro selection of miltefosine resistance in promastigotes of *Leishmania donovani* from Nepal: genomic and metabolomic characterization. *Mol Microbiol* **99**:1134–1148. doi:10.1111/mmi.13291

Simpson AGB, Stevens JR, Lukeš J. 2006. The evolution and diversity of kinetoplastid flagellates. *Trends Parasitol* **22**:168–174. doi:10.1016/j.pt.2006.02.006

Singh R, Kumar D, Duncan RC, Nakhasi HL, Salotra P. 2010. Overexpression of histone H2A modulates drug susceptibility in Leishmania parasites. *Int J Antimicrob Agents* **36**:50–57. doi:10.1016/j.ijantimicag.2010.03.012

1345 Sinha R, C MM, Raghwan, Das Subhadeep, Das Sonali, Shadab M, Chowdhury R, Tripathy S,
1346 Ali N. 2018. Genome Plasticity in Cultured Leishmania donovani: Comparison of Early
1347 and Late Passages. *Front Microbiol* **9**. doi:10.3389/fmicb.2018.01279
1348 Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature*
1349 **415**:1022–1024. doi:10.1038/4151022a
1350 Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with
1351 thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690.
1352 doi:10.1093/bioinformatics/btl446
1353 Sterkers Y, Crobu L, Lachaud L, Pagès M, Bastien P. 2014. Parasexuality and mosaic
1354 aneuploidy in *Leishmania*: alternative genetics. *Trends Parasitol* **30**:429–435.
1355 doi:10.1016/j.pt.2014.07.002
1356 Sterkers Y, Lachaud L, Crobu L, Bastien P, Pagès M. 2011. FISH analysis reveals aneuploidy
1357 and continual generation of chromosomal mosaicism in *Leishmania major*. *Cell*
1358 *Microbiol* **13**:274–283. doi:10.1111/j.1462-5822.2010.01534.x
1359 Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. *Mol Biol Evol* **28**:63–70.
1360 doi:10.1093/molbev/msq249
1361 Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarizes and Visualizes Long Lists
1362 of Gene Ontology Terms. *PLOS ONE* **6**:e21800. doi:10.1371/journal.pone.0021800
1363 Swenerton RK, Knudsen GM, Sajid M, Kelly BL, McKerrow JH. 2010. *Leishmania* subtilisin is a
1364 maturase for the trypanothione reductase system and contributes to disease
1365 pathology. *J Biol Chem* **285**:31120–31129. doi:10.1074/jbc.M110.114462
1366 Teixeira DG, Monteiro GRG, Martins DRA, Fernandes MZ, Macedo-Silva V, Ansaldi M,
1367 Nascimento PRP, Kurtz MA, Streit JA, Ximenes MFFM, Pearson RD, Miles A, Blackwell
1368 JM, Wilson ME, Kitchen A, Donelson JE, Lima JPMS, Jeronimo SMB. 2017.
1369 Comparative analyses of whole genome sequences of *Leishmania infantum* isolates
1370 from humans and dogs in northeastern Brazil. *Int J Parasitol* **47**:655–665.
1371 doi:10.1016/j.ijpara.2017.04.004
1372 Thakur L, Singh KK, Shanker V, Negi A, Jain A, Matlashewski G, Jain M. 2018. Atypical
1373 leishmaniasis: A global perspective with emphasis on the Indian subcontinent. *PLoS*
1374 *Negl Trop Dis* **12**:e0006659. doi:10.1371/journal.pntd.0006659
1375 Uzcategui NL, Zhou Y, Figarella K, Ye J, Mukhopadhyay R, Bhattacharjee H. 2008. Alteration
1376 in glycerol and metalloid permeability by a single mutation in the extracellular C-loop
1377 of *Leishmania major* aquaglyceroporin LmAQP1. *Mol Microbiol* **70**:1477–1486.
1378 doi:10.1111/j.1365-2958.2008.06494.x
1379 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan
1380 T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S,
1381 DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome
1382 Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma Ed Board Andreas*
1383 *Baxevanis Al* **11**:11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43
1384 Weir W, Capewell P, Foth B, Clucas C, Pountain A, Steketee P, Veitch N, Koffi M, De Meeûs T,
1385 Kaboré J, Camara M, Cooper A, Tait A, Jamonneau V, Bucheton B, Berriman M,
1386 MacLeod A. 2016. Population genomics reveals the origin and asexual evolution of
1387 human infective trypanosomes. *eLife* **5**:e11473. doi:10.7554/eLife.11473
1388 Wright S. 1938. The Distribution of Gene Frequencies Under Irreversible Mutation. *Proc Natl*
1389 *Acad Sci U S A* **24**:253–259.
1390 Zackay A, Cotton JA, Sanders M, Hailu A, Nasereddin A, Warburg A, Jaffe CL. 2018. Genome
1391 wide comparison of Ethiopian *Leishmania donovani* strains reveals differences

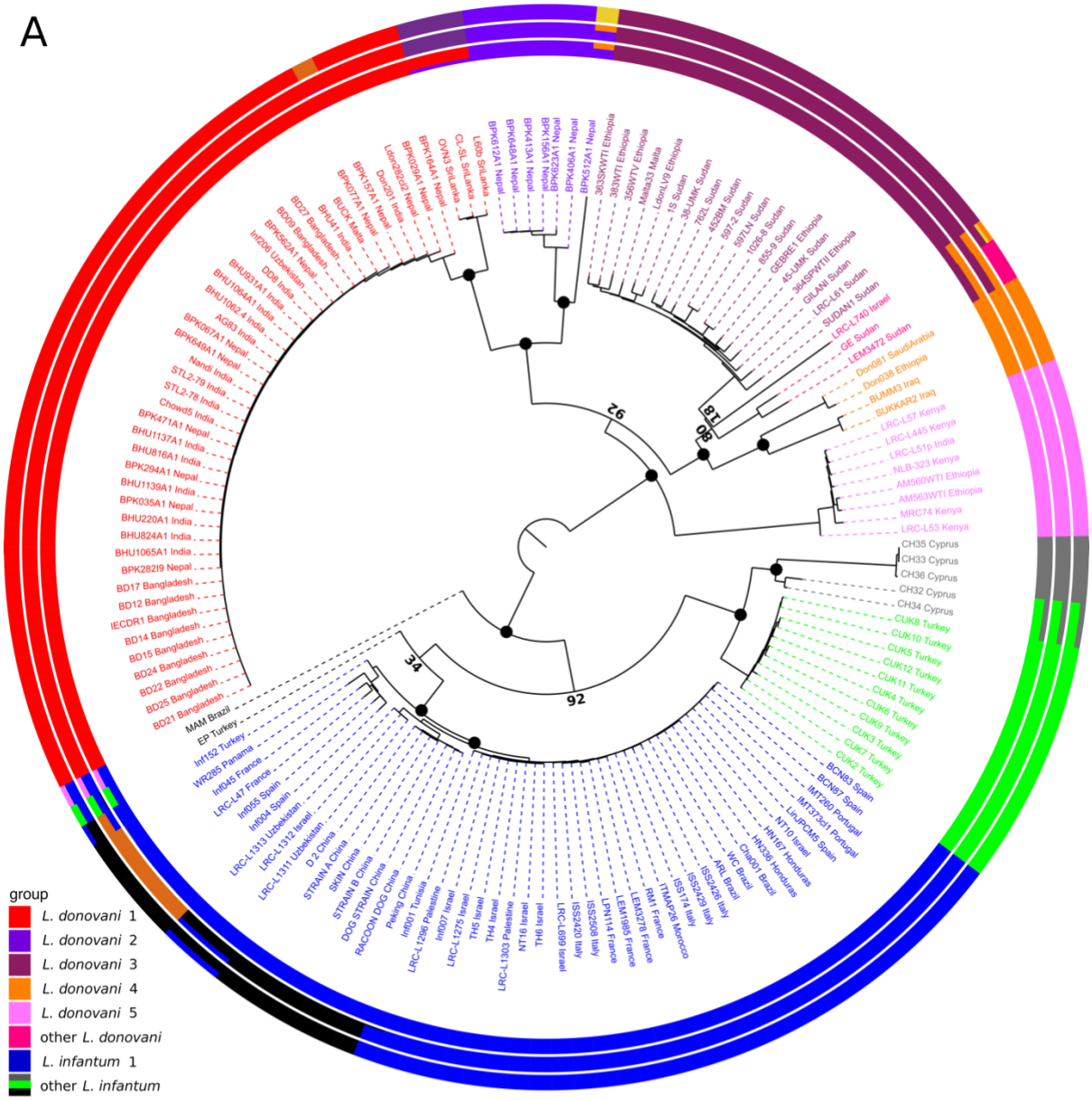
potentially related to parasite survival. *PLOS Genet* **14**:e1007133. doi:10.1371/journal.pgen.1007133

Zhang C-Y, Lu X-J, Du X-Q, Jian J, Shu L, Ma Y. 2013. Phylogenetic and Evolutionary Analysis of Chinese *Leishmania* Isolates Based on Multilocus Sequence Typing. *PLoS ONE* **8**. doi:10.1371/journal.pone.0063124

Zhang WW, Ramasamy G, McCall L-I, Haydock A, Ranasinghe S, Abeygunasekara P, Sirimanna G, Wickremasinghe R, Myler P, Matlashewski G. 2014. Genetic Analysis of *Leishmania donovani* Tropism Using a Naturally Attenuated Cutaneous Strain. *PLOS Pathog* **10**:e1004244. doi:10.1371/journal.ppat.1004244

Zijlstra E, Musa A, Khalil E, El Hassan I, El-Hassan A. 2003. Post-kala-azar dermal leishmaniasis. *Lancet Infect Dis* **3**:87–98. doi:10.1016/S1473-3099(03)00517-6

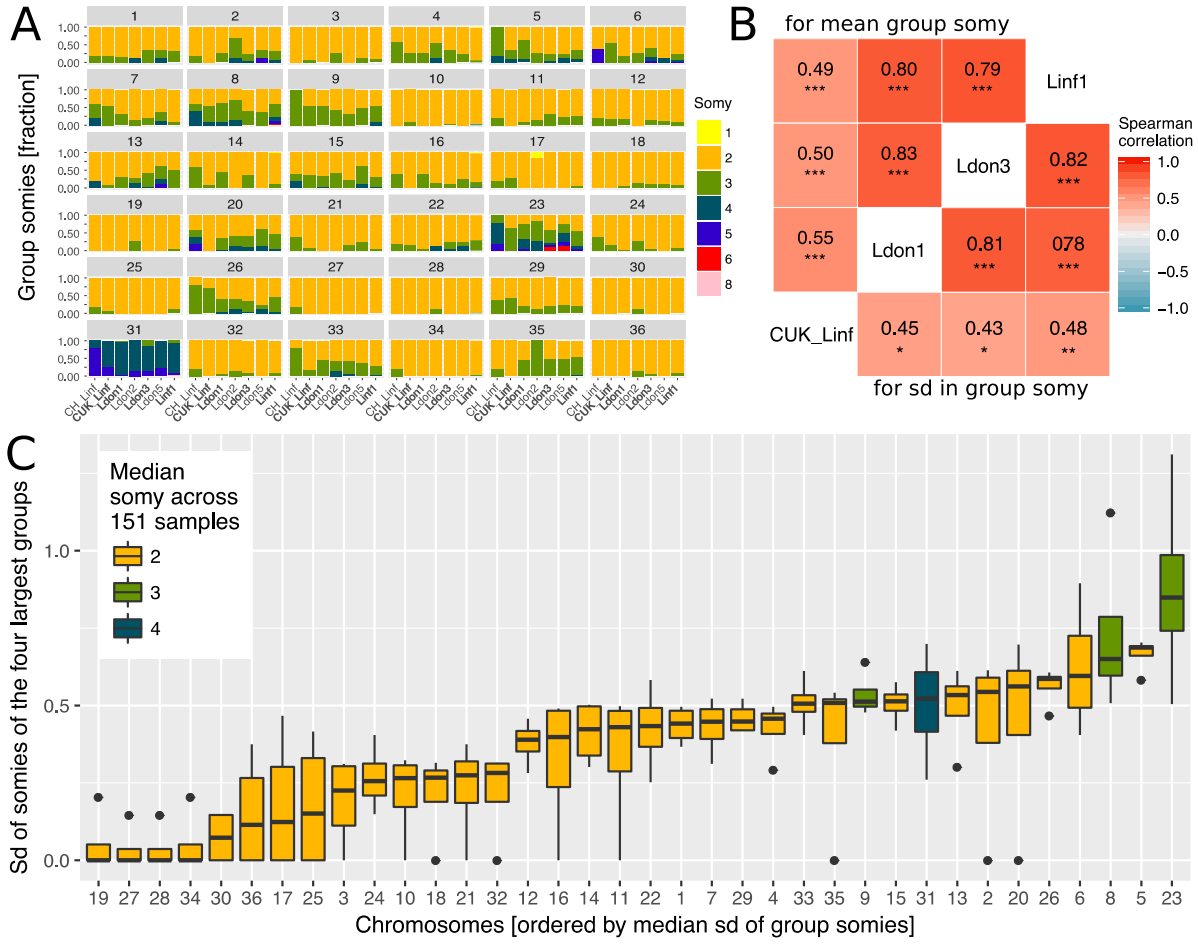
1404 Figure 1: Sample phylogeny and distribution



1405

Fig. 1: Sample phylogeny and distribution. A) Phylogeny of all 151 samples of the *L. donovani* complex. The phylogeny was calculated with neighbour joining based on Nei's distances and rooted based on the inclusion of isolates of *L. mexicana* (U1103.v1), *L. tropica* (P283) and *L. major* (LmjFried) (outgroups not shown in the phylogeny). Bootstrap values are shown for prominent nodes in the phylogeny as black circles for values of 100 and otherwise the respective support value. The groupings shown in the outer circles were calculated by admixture with *K*=8, *K*=11 and *K*=13 (see Material & Methods). Groups labelled with different colours were defined based on the phylogeny and include monophyletic groups as well as groups that are polyphyletic and/ or largely influenced by hybridisation (indicated by "other"). B) Map of the sampling locations. Groups are indicated by the different colours. Sample sizes by country of origin are visualised by the sizes of the circles.

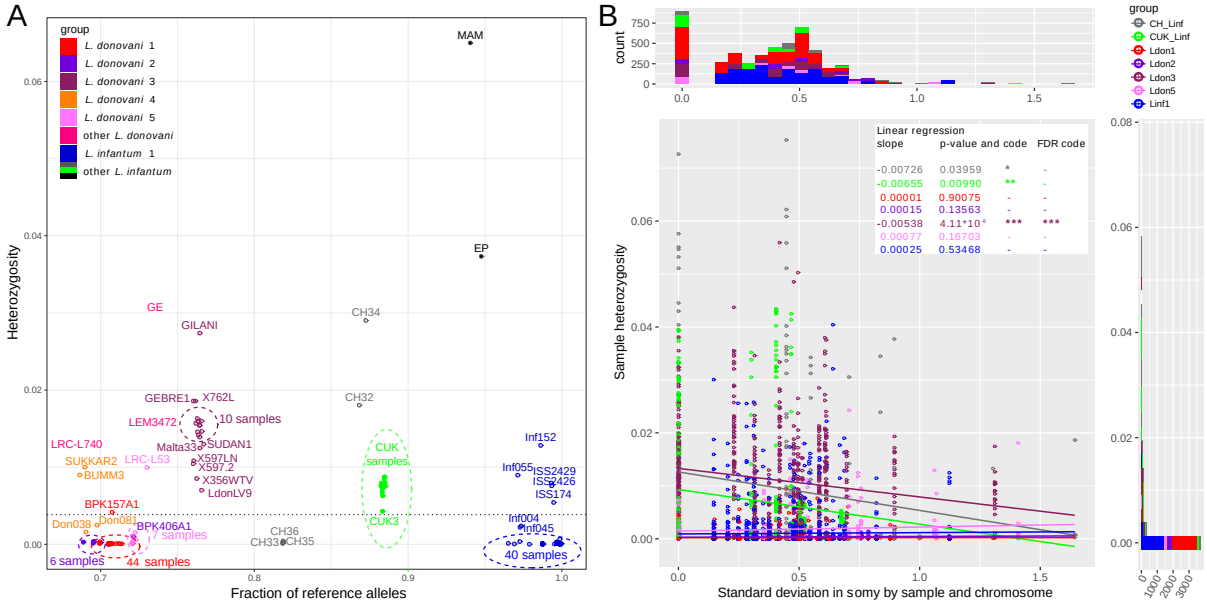
Figure 2: Chromosome-specific somy variability



Chromosome-specific somy variability. A) Somy variability is displayed for the 7 largest groups (≥5 isolates) for each chromosome as fractions of isolates with the respective somies. The four largest groups (≥9 samples per group) are indicated in bold. B) The heatmap shows the Spearman correlations of chromosome-specific somy statistics between the four largest groups, measured as the mean group somies (upper triangle) and the standard deviation (sd) of chromosome somies (lower triangle), respectively. False discovery rates (FDR) of each correlation are indicated by asterisks (*: <0.05, **: <0.01, ***: <0.001). C) Boxplots show the distribution of variability in chromosome-specific somy across the four largest groups used as independent replicates across the species range. Medians estimate the chromosome-specific variation in somy.

1429 Figure 3: Whole genome sample heterozygosities

1430

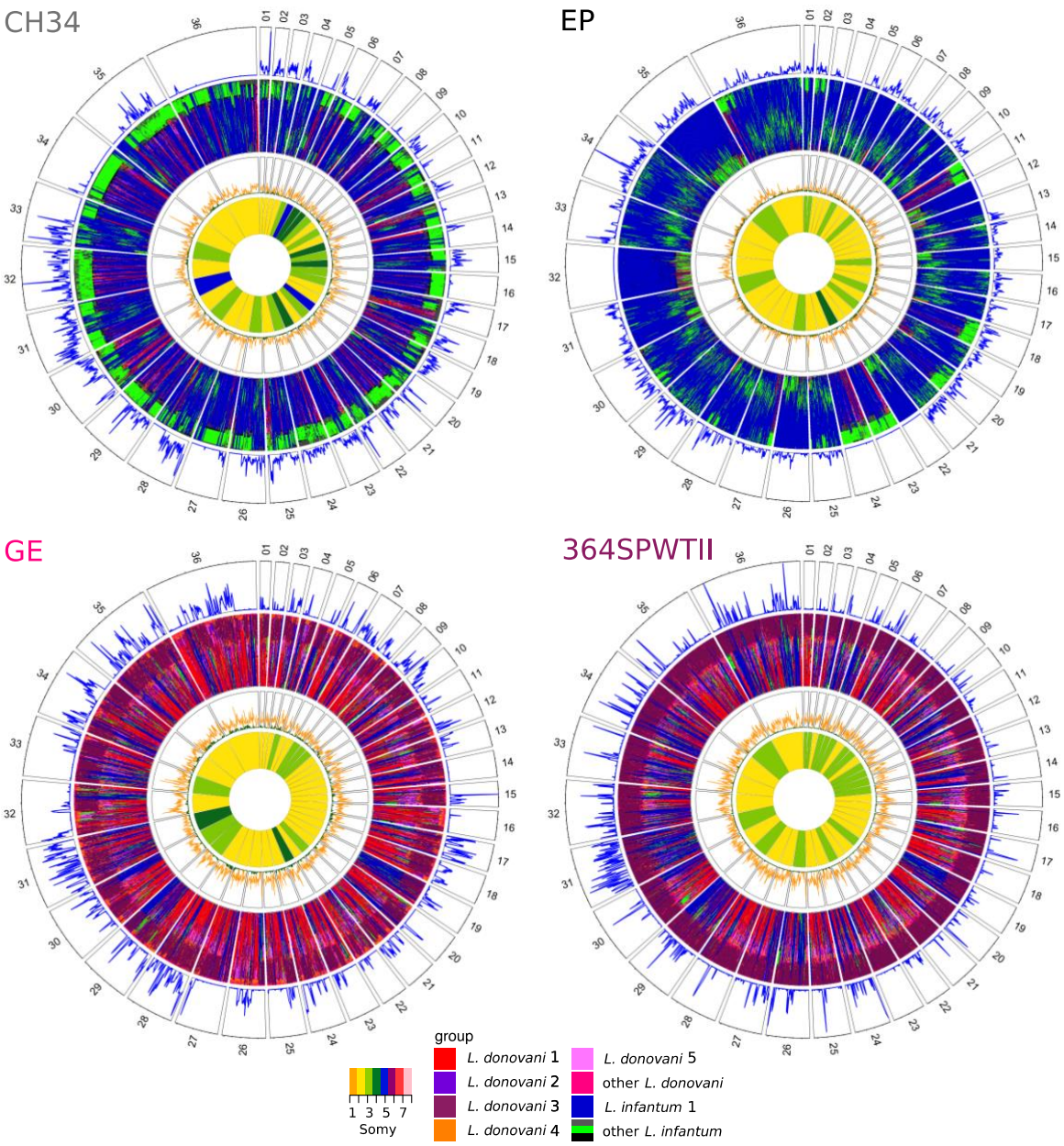


1431

1432

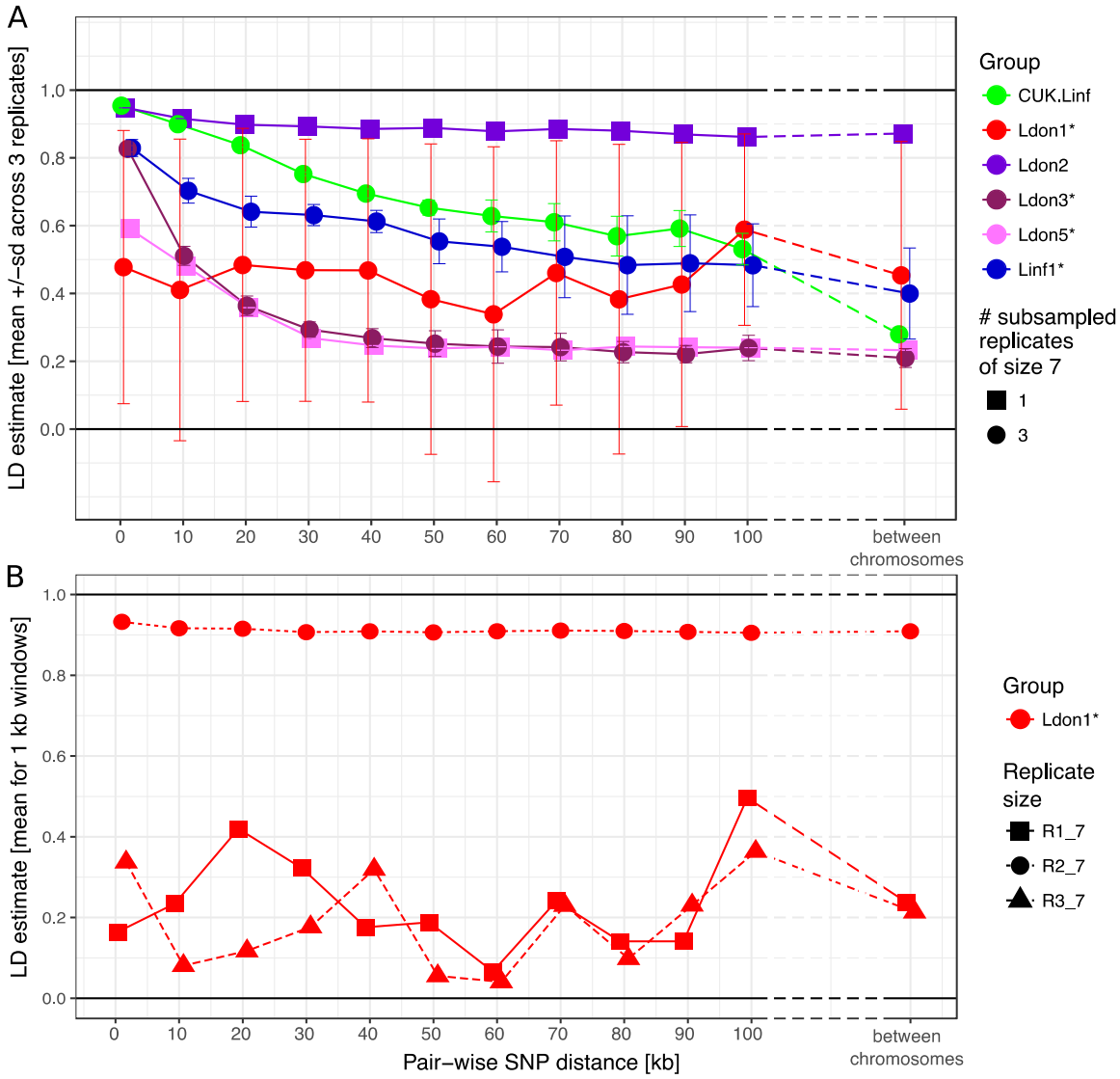
1433 Whole genome sample heterozygosities. A) Whole genome heterozygosities versus fraction of
1434 reference alleles. The fraction of reference alleles is calculated across all 395,602 SNP loci in the data
1435 set. Isolate names are written unless they are present in dense clusters indicated by dashed-line
1436 circles. Groups are indicated by colour as defined in figure 1. The dashed horizontal line at a genome-
1437 wide heterozygosity of 0.004 was chosen to separate samples with putative recent between-strain
1438 hybridisation history. B) Relationship between chromosome-specific somy variability and sample
1439 heterozygosity. The scatterplot describes the relationship between the standard deviation in
1440 chromosome-specific somy by group (groups with ≥5 samples) against the chromosome-specific
1441 sample heterozygosity. Linear regressions were performed for each group. Asterisks indicate
1442 statistical significance of the estimated regression slope with *: <0.05, **: <0.01, ***: <0.001 or '-' for
1443 not significant. Marginal histograms on the top and on the right correspond to the x-values and the y-
1444 values of the scatterplot, respectively. Groups are indicated by the different colours.

Figure 4: Window based analysis of relatedness



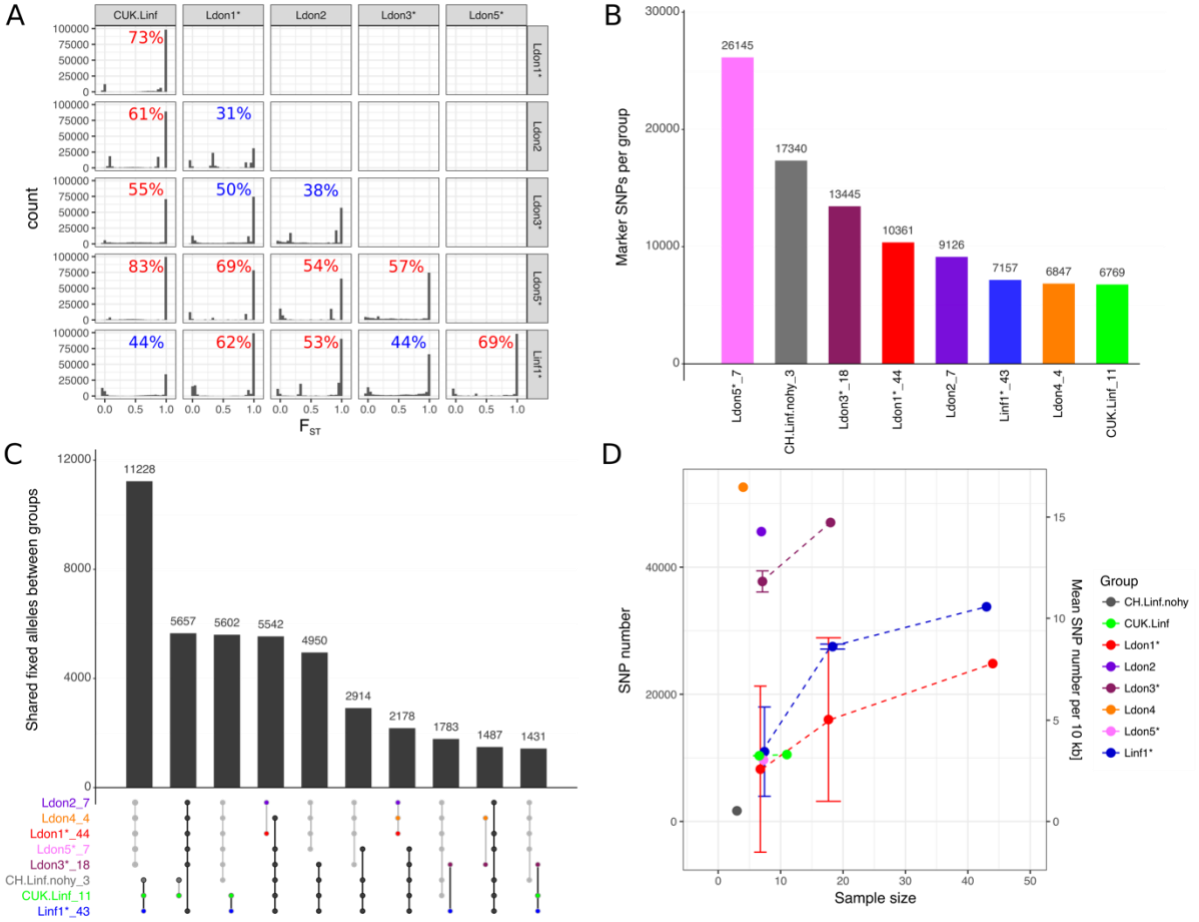
Window-based analysis of relatedness. Each circos plot shows four different genomic features of the isolate named in each top left corner. In the four different rings, pies correspond to the different chromosomes labelled by chromosome number. The three outer rings show a window-based analysis for a window size of 10 kb. Starting from the outer ring, they show: 1. Heterozygosity with the number of heterozygous sites ranging from 0 to 98, 146, 90 and 85 sites per window for CH34, EP, GE and 364SPWTII, respectively, 2. A heatmap coloured by groups of the 60 genetically closest isolates based on Nei's D and starting with the closest sample at the outer margin and the 60th furthest isolate at the inner margin, 3. Nei's D to the closest (green) and the 60th closest isolate (orange) scaled from 0 to 1. The innermost circle shows the colour-coded somy.

Figure 5: LD decay with genomic distance



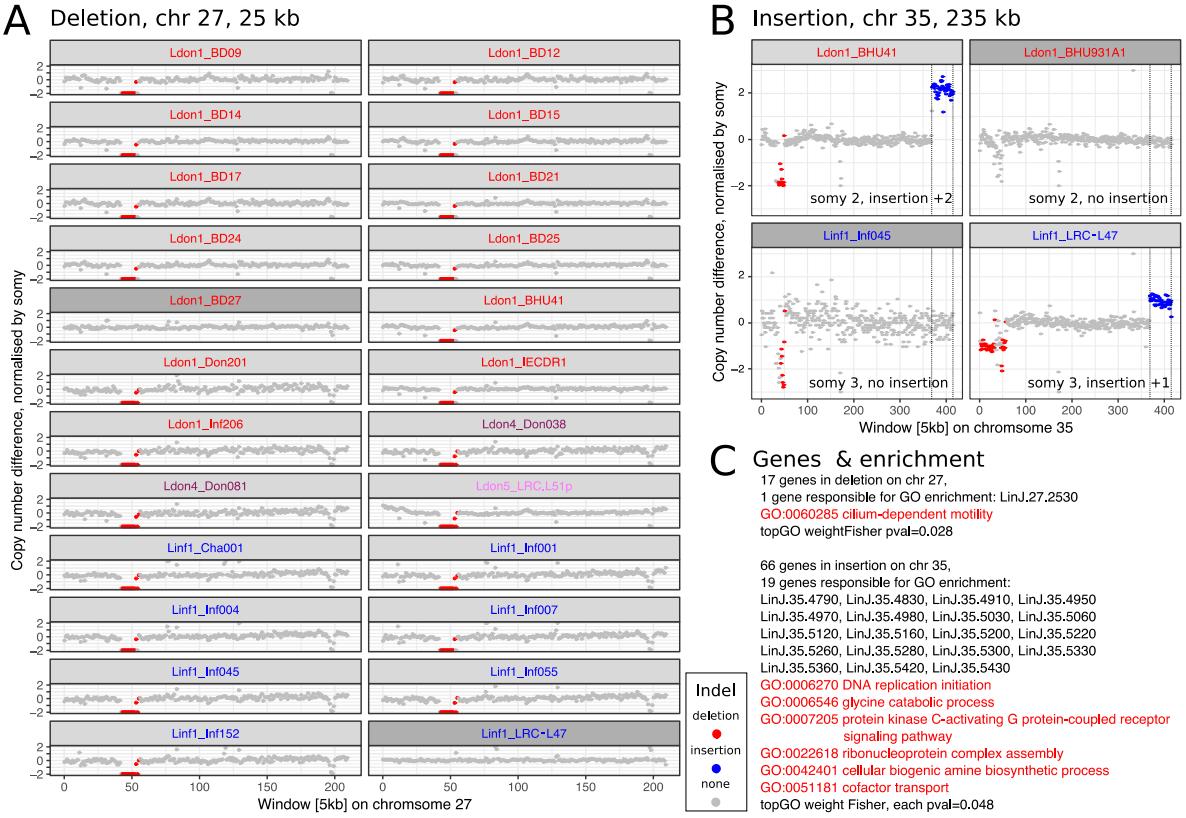
LD decay with genomic distance. A) LD decay was measured for the six largest groups removing isolates that were identified as putative strain mixtures (indicated by *; see Material & Methods). Groups with more than seven isolates per group were sub-sampled to pseudo-replicates of seven isolates per group three times to make LD estimates comparable between groups. Mean and standard deviation across the three pseudo-replicates are shown where applicable. B) LD decay with distance is shown for the three pseudo-replicates for the Ldon1 group. A and B) Data for individual replicates was calculated as means of 1kb windows for SNP pairs of the stated genomic distance. For LD estimates between chromosomes, 100 SNPs were randomly sampled per chromosome and means across all pair-wise combinations between chromosomes are shown. This procedure was done twice independently but as differences between both such replicates were negligible, only the results of one replicate are shown.

Figure 6: Differentiated and segregating SNPs between and within groups



Differentiated and segregating SNPs between and within groups. For this analysis isolates that were shown to be mixtures of clones or hybrids between groups were removed (indicated by '*', see also Material & Methods). Groups sizes after removal of those isolates are specified in panels A and C. A) F_{ST} values between pairwise group comparisons. The fraction of differentially fixed SNPs (F_{ST}=1) for each pairwise group comparison is indicated at the top right corner of each plot. Percentages larger than 50% are coloured in red, otherwise blue. B) The number of marker SNPs for each group, i.e. SNPs that are differentially fixed in one group versus all others. C) Number of SNPs that are differentially fixed between sets of groups. Groups fixed for the same allele are indicated in the bottom panel through connecting points corresponding to the specific groups. Grey and black lines connect sets of groups monomorphic for the alternate and reference allele, respectively. D) Number and density of SNPs segregating in the respective groups. As sample sizes of the different groups vary, figures are also shown for three random sub-samples of the larger groups. Results of sub-sampling are displayed as mean and sd.

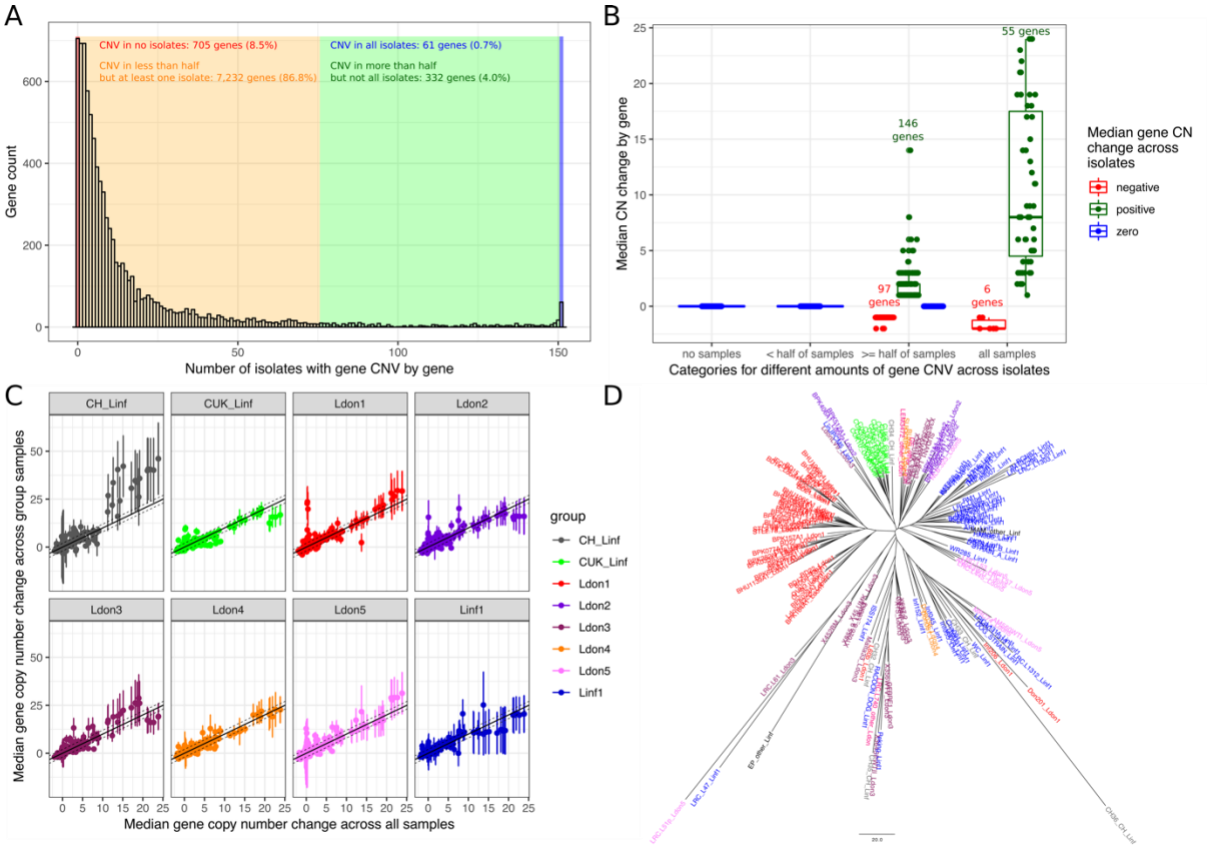
1490 Figure 7: Large CNVs that are shared between both species.
1491



1492
1493

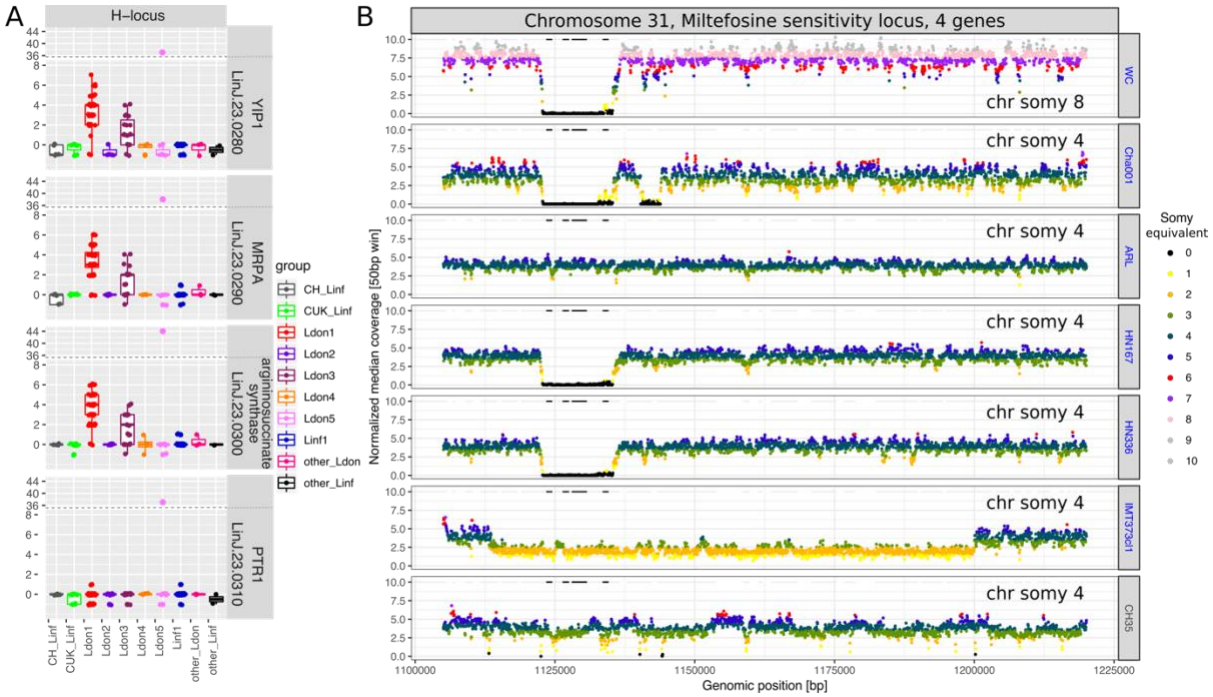
1494 Two large CNVs that are shared between both species. A) Chromosome 27 has a 25 kb long deletion
1495 that is present in 15% of all samples and four different groups. All chromosomes 27 that have this
1496 deletion in our dataset are diploid and the deletion results in a loss of this allele in the respective
1497 sample. B) The duplication on chromosome 35 is 235 kb long and present in one isolate of group Ldon1
1498 and Linf1, respectively. The insertion is once present on a disomic background with a 2-fold increase
1499 and once on a trisomic background with a 1-fold increase. For A) and B) a few closely related samples
1500 not harbouring the respective CNV are also displayed and highlighted in dark grey. Group identities
1501 are indicated by colours of the isolate name. C) Genes present in the respective CNV along with GO
1502 enrichment results using topGO (Alexa et al., 2006). Details on both CNVs can be found in table S6:
1503 unique CNVs with ids 150 and 215, respectively. The CNV characterisation of the corresponding
1504 isolates can be found in table S5.

Figure 8: Gene copy number variation across groups



Gene copy number variation across groups. A) CN abundances by gene across all 151 isolates. Genes are grouped in four categories (identified by different colours) depending on how many isolates are affected by CN variation in the respective gene. B) Median copy number changes for each gene are shown (individual dots) and summarised for the four different categories also used in sub-figure A including the direction of effect sizes using boxplots. C) Correlations of the median gene copy number across all samples and each respective phylogenetic group. D) Neighbour joining tree using gene CN profiles for each sample.

Figure 9: Copy number variation of putative drug resistance genes



Copy number variation of putative drug resistance genes. A) Copy numbers (CNs) for all four genes on the H-locus are shown for all 151 samples across all 10 different (sub-)groups. B) Genome coverage in the genomic regions surrounding the MSL in all six samples showing a deletion and one sample with no CN reduction. Genome coverage for 50 kb windows is normalised by the haploid chromosome coverage and colours indicate the somy equivalent coverage of the respective window. The genes, LinJ.31.2370, LinJ.31.2380, LinJ.31.2390 and LinJ.31.2400, are marked as black horizontal lines. Colours of the sample names indicate group colours used throughout this study.

1527 Table 1: Summary of the hybrid analysis.

Category	ID	Description	Interpretation	# Samples	Fraction of samples	Sample identities
Initial definition of the 53 (35%) putative hybrids	A1	"High" genome-wide heterozygosity (>=0.004)	initial indicator for putative hybrids	46	30%	BPK157A1, BUMM3, CH32, CH34, CUK10, CUK11, CUK12, CUK2, CUK3, CUK4, CUK5, CUK6, CUK7, CUK8, CUK9, EP, GE, GEBRE1, GILANI, Inf055, Inf152, ISS174, ISS2426, ISS2429, LdonLV9, LEM3472, LRC-L53, LRC-L61, LRC-L740, Malta33, MAM, SUDAN1, SUKKAR2, 1026-8, 1S, 356WTV, 363SKWTI, 364SPWTII, 38-UMK, 383WTI, 45-UMK, 452BM, 597-2, 597LN, 762L, 855-9
	A2	“Admixed” between groups (admixture analysis)	initial indicator for putative hybrids	15	10%	BPK512A1, CH32, CH34, CL-SL, EP, GE, Inf152, L60b, LEM3472, LRC-L1311, LRC-L1312, LRC-L1313, LRC-L740, MAM, OVN3
Detailed investigation of the 53 (35%) putative hybrids	B1	Heterozygous sites distributed relatively evenly across the genome and allele frequency profiles match coverage based somy estimates	putative patterns of sexual crossing (F1 / F2+), however, cannot be verified without identified putative parents; alternative explanation could be new mutations that are dominating the sample population through a recent bottleneck (e.g. cloning)	18	12%	Inf055, GEBRE1, LdonLV9, LRC.L61, SUDAN1, 1026-8, 1S, 356WTV, 363SKWTI, 364SPWTII, 38-UMK, 383WTI, 45-UMK, 452BM, 597-2, 597LN, 762L, 855-9
	B2	Evidence for parents between different groups (or between 2 distinct strains as previously shown for the CUK samples) alternating in the genome in a block like pattern	putative patterns of sexual crossing (F2+), i.e. "hybrids"	16 (+1)	10% (11%)	CH32, CH34, CUK10, CUK11, CUK12, CUK2, CUK3, CUK4, CUK5, CUK6, CUK7, CUK8, CUK9, EP, GE, LEM3472, (LRC-L740)
	B3	Extreme allele frequency variants only	mixture of two different high versus low frequency clones or low frequency new mutations distributed across haplotypes in the sample	7	5%	BPK157A1, Inf152, ISS174, ISS2426, ISS2429, LRC-L53, MAM
	B4	Intermediate peak allele frequency distributions including extreme frequency peaks	mixture of scenarios B1 and B3, i.e. as B3 but high frequency clone has heterozygous sites itself	4	3%	BUMM3, LRC-L740, Malta33, SUKKAR2
	B5	no clear peak pattern of allele frequencies (several peaks at atypical frequencies)	mixture of several clones	1	0.01%	GILANI
	B6	to few heterozygous sites present to draw further conclusions beyond admixture results	signatures are shadowed by too little segregating variation	7	5%	BPK512A1, CL-SL, L60b, LRC-L1311, LRC-L1312, LRC-L1313, OVN3

Table 2: Summary of genetic variation across 151 isolates of the *L. donovani* complex for previously described loci involved in resistance or treatment failure of antimonial drugs and Miltefosine.

locus / complex	gene id			gene name	function prediction	involved in resistance (R) / treatment failure (TF) to drug:	reference	evidence from reference	gene copy number (gene CN)
	***L. infantum*, JPCM5, v41**	***L. infantum*, JPCM5, v38**	***L. donovani* ortholog, BPK282A1, v41**						
H-locus	LINF_230007700	LinJ.23.0280	LdBPK_230280	terbinafine resistance gene (HTBF), (YIP1)		Antimonials (R)	Callahan and Beverley, 1991; Dias et al., 2007	The *Leishmania* H region is frequently amplified in drug-resistant lines and is associated with metal resistance (genes YIP1, MRPA, PTR1).	Genes have an increased CN in 30% (CN +1 to +44), and reduced CN in 9% (CN -1). 37% of all samples have an insertion including at least 3 genes (always YIP1, MRPA and argininosuccinate synthase). These amplifications are in groups Ldon1 (42/45), Ldon3 (13/19) and Ldon5 (1/8). The insertion boundaries in isolates from groups Ldon1 and Ldon3 are shared (Fig. S24 A).
	LINF_230007800	LinJ.23.0290	LdBPK_230290	P-glycoprotein A (MRPA); pentamidine resistance protein 1	ATP-binding cassette (ABC) transporter, ABC-thiol transporter	Antimonials (R)	Callahan and Beverley, 1991; Dias et al., 2007; Leprohon et al., 2009	Increased expression of MRPA is often due to the amplification of its gene in antimony-resistant strains.	
	LINF_230007900	LinJ.23.0300	LdBPK_230300		argininosuccinate synthase - putative	Antimonials	Grondin et al., 1993; Leprohon et al., 2009		
	LINF_230008000	LinJ.23.0310	LdBPK_230310	Pteridine reductase 1 (PTR1)		Antimonials (R)	Callahan and Beverley, 1991; Dias et al., 2007	see above	
Mitogen-activated protein kinase, MAPK1	LINF_360076200	LinJ.36.6760	LdBPK_366760	LMPK, mitogen-activated protein kinase	protein phosphorylation	Antimonials (R)	Singh et al., 2010; Ashutosh et al., 2012	Conflicting evidence between up- and down-regulation of Mitogen-Activated Protein Kinase 1 between different studies.	45% of all isolates showed an increased CN, with all isolates of Ldon1 andLdon3 being affected and smaller fractions in other *L. donovani* groups (Fig. S24 A).
Aqua-glyceroporin, AQP1	LINF_310005100	LinJ.31.0030	LdBPK_310030	Aquaglyceroporin 1, AQP1	drug transmembrane transport	Antimonials (R)	Gourbal et al., 2004; Uzcategui et al., 2008; Monte-Neto et al., 2015; Andrade et al., 2016; Imamura et al., 2016	A frequently resistant *L. donovani* population has a two base-pair insertion in AQP1 preventing antimonial transport. Increased resistance with decrease in gene CN or expression, while increase leads to higher drug sensitivity.	Gene CN deletions and insertions of small effect sizes (CN -2 to -1 and +1 to +3) are present in 6% and 35% of isolates but never leading to loss of the locus.

locus / complex	gene id			gene name	function prediction	involved in resistance (R) / treatment failure (TF) to drug:	reference	evidence from reference	gene copy number (gene CN)
	L. infantum, JPCM5, v41	*L. infantum*, JPCM5, v38	*L. donovani* ortholog, BPK282A1, v41						
Miltefosine transporter and associated genes	LINF_130020800	LinJ.13.1590	LdBPK_131590	Miltefosine transporter, LdMT	phospholipid transport	Miltefosine (R)	Pérez-Victoria et al., 2006; Shaw et al., 2016	Gene deletion or different changes in two different strains evolved in promastigote culture for Miltefosine resistance. strain Sb-S: locus deletion and A691P; strain Sb-R: E197D	15 isolates: +1 gene CNV (CUK, Lon1, Ldon2, Ldon3, Ldon5)
	LINF_130020900	LinJ.13.1600	LdBPK_131600	hypothetical protein	unknown function	Miltefosine (R)	Shaw et al., 2016	Deleted along with the Miltefosine transporter gene in a single line evolved for Miltefosine resistance in promastigote culture.	3 isolates: +1 gene CNV (Ldon1, Linf1)
	LINF_320015500	LinJ.32.1040	LdBPK_321040	Ros3, LdRos3	Vps23 core domain containing protein - putative	Miltefosine (R)	Pérez-Victoria et al., 2006	Putative subunit of LdMT; LdMT and LdRos3 seem to form part of the same translocation machinery that determines flippase activity and Miltefosine sensitivity in *Leishmania*.	1 isolate: +1 gene CNV (Ldon1)
Miltefosine sensitivity locus, MSL	LINF_310031200	LinJ.31.2370	LdBPK_312380		3'-nucleotidase/ nuclease - putative	Miltefosine (TF)	Carnielli et al., 2018	MSL: a deletion of this locus was associated with Miltefosine treatment failure in Brazil. While the frequency of the MSL was still relatively high in the North-East it was almost absent in the South-East of Brazil, and it was absent in *L.infantum*/ *L.donovani* in the Old World.	Genes have a reduced CN in 55% (CN -1 to -8) and increased in 4% (CN +1). 4 isolates, show a complete loss of the MSL at identical boundaries: WC, Cha001, HN167 and HN336 (2/4 isolates from Brazil, 2/2 isolates from Honduras). 2 isolates show a reduction of all 4 genes at this locus but with various deletion boundaries: IMT373cl1 (Portugal), CH35 (Cyprus) (Fig. 9 B).
	LINF_310031300	LinJ.31.2380	LdBPK_312380		3'-nucleotidase/ nuclease - putative	Miltefosine (TF)	Carnielli et al., 2018		
	LINF_310031400	LinJ.31.2390	LdBPK_312390		helicase-like protein	Miltefosine (TF)	Carnielli et al., 2018		
	LINF_310031500	LinJ.31.2400	LdBPK_312320, LdBPK_312400		3 -2-trans-enoyl-CoA isomerase - mitochondrial precursor - putative	Miltefosine (TF)	Carnielli et al., 2018		

Table 3: Candidate genes putatively involved in pathogenesis associated differences between *L. donovani* and *L. infantum*. Candidates were identified through GO enrichment analysis of moderate to high effect variants between both species across our 151 isolates.

Gene name	**Gene codes v41 (v38) Tritryp (http://tritrypdb.org/tritrypdb/)**	**Annotation**	**Fixed genomic variation between *L. infantum* and *L. donovani* (changes *L.inf* > *L.don*)**	**Evidence for pathogenic function**
Tir chaperone protein	LINF_040012200 (LinJ.04.0710), LINF_340038600 (LinJ.34.2950)	Tir chaperone protein (CesT) family/PDZ domain containing protein - putative, Tir chaperone protein (CesT) family - putative	nt 362A>G; aa Glu121Gly nt 594A>G aa Gln198Gln nt 1659A>C; aa Lys553Asn nt 1703A>G; aa Asn568Ser	Part of secretion system to deliver virulence effector proteins into the host cell cytosol in gram-negative bacteria; secreted proteins require chaperones to maintain function (Delahay et al., 2002).
Subtilisin protease	LINF_130015300 (LinJ.13.0940 and LinJ.13.0930, -strand, are fused in v41 with an extra 54 bp in between them)	subtilisin-like serine peptidase	nt 2813T>G; aa Phe938Cys nt 3346G>A; aa Gly1116Ser nt 4389G>A; aa Pro1463Pro* nt 5014A>C; aa Ser1672Arg*	Shown to be essential for full virulence and involved in detoxification of ROS in *L. donovani* (Swenerton et al., 2010).
Bardet-biedl syndrome 1 protein	LINF_350047600 (LinJ.35.4250)	Bardet-Biedl syndrome 1 protein homolog (BBS1-like protein 1) - putative	nt 531C>T; aa Ser177Ser nt 580G>A; aa Ala194Thr nt 1038C>A; aa Arg346Arg nt 1221T>C; aa Gly407Gly nt 1310C>T; aa Ala437Val	*Leishmania* BBS1 know-out mutants have reduced infectivity for *in vivo* macrophages and infection of BALB/c mice was severely compromised (Price et al., 2013).

*Nucleotide (nt) and amino acid (aa) changes in LinJ.13.0930 (v38) have been adapted to positions to its fused version LINF_130015300 (v41) in this table. Positions for v38 can be found in table S3.